

<https://doi.org/10.1038/s41746-025-01979-3>

Personalised modelling of routine variability and affective states

Adrien Choi , Danielle Lottridge & Jim Warren

Multimodal smartphone sensor data provide rich insights into real-world behavioural patterns associated with anxiety and depression symptoms. We propose that variability in daily routines across life aspects may act as a personalised digital marker, linking behaviours to self-reported affective states. Using non-negative matrix factorisation (NMF), we decomposed mobile sensing data into individual-specific routines and their weekly variability. Generalised linear models (GLMs) were built per individual to associate variability from specific sensing categories with anxiety or depression states. Using a large language model (LLM; GPT-4o), we translated the modelling results into more accessible language that potentially helps individuals engage with their own routine phenotypes for self-regulation insights. Population-level grouping of GLMs revealed significant between-group differences in mental health measures. The LLM output shows a potential to uncover personalised behavioural patterns and support self-understanding of mood-related drivers. This approach may also inform group-based interventions by identifying individuals likely to benefit from shared treatment strategies based on routine similarity.

Compared to the relatively established neurobiological basis of depression and anxiety, considerable challenges have been presented to understand real-world behavioural manifestations of these symptoms of varying degrees¹. Existing research has primarily relied on questionnaires and interviews, which require participants to recall past behaviours^{2–5}. Although questionnaires provide a certain level of quantitative capability, they are prone to various types of information bias⁶, making them less reliable for studying long-term human behaviours that typically need to be observed from a third-party perspective. Multimodal sensor data captured on personal smartphones and other connected devices are a fruitful source of granular behavioural data^{7,8}. When recorded over the long term, behavioural data exhibits inherent structural and repetitive properties⁹. This presents a research opportunity to assess the extent to which individuals establish specific routine patterns across various aspects of life and how these patterns are associated with mental health outcomes.

A routine refers to a sequence of behaviours or activities that a person performs regularly to maintain daily life. Primary routines¹⁰, such as sleep-wake cycles, are governed by intrinsic circadian clocks, with inputs also provided by zeitgebers¹¹, which are external environmental and social cues. Secondary routines¹⁰, which may include activities like exercise, mealtimes, or work schedules, are more flexible and often shaped by societal norms and individual choices, but they still interact with primary routines and zeitgebers. For instance, social obligations or work demands may alter an individual's exercise or eating patterns, which, in turn, can affect sleep and overall circadian stability. Routine disruptions have been linked to the

severity of anxiety and depression symptoms^{12,13}; however, little is known about the long-term effects of specific types of routines¹², as they can only be revealed with extensive logs of behavioural data. Furthermore, such data encodes individual-level digital phenotypes (DP)¹⁴, which can serve as the fundamental components of a dynamic virtual representation of a patient, the digital twins (DT)¹⁵. These digital twins faithfully mirror an individual's real-world behaviours with analytic and predictive capabilities. The intrinsic links between DP and real-world human behaviours are bidirectional. From an evolutionary perspective, the extended phenotype involves organisms altering their environment to enhance fitness, while the digital phenotype reflects human modifications of the online environment to exercise species-typical capabilities and derive benefits from digital information¹⁶. While an end-to-end pipeline from captured sensor data to a clinical endpoint could provide predictions for early intervention, a mechanistic approach that emphasises studying intermediate outcomes and underlying disease mechanisms might also advance the field by uncovering the interactions between higher-level behavioural, psychological, and biological states^{8,17,18}.

In machine learning, such intermediate-level processing often manifests during the stage of feature engineering. Most studies that modelled the relationship between behavioural data and mental health outcomes aggregated the data into time-windows without capturing temporal change^{19–25}. Fewer studies looked at how changes in behaviours associated with mental health measures^{26–29}. For example, one study²⁹ aggregated behavioural data based on metrics of depression symptoms, including changes in sleep, diminished ability to concentrate, diminished interest in activities, low

mood, and energy. Several metrics were found to be associated with self-reported depression scores, and the study achieved a weekly prediction recall of 81.5% and a precision of 69.1%. Another study²⁷ applied association rule mining to contextually filter rules that could distinguish between depression and non-depression groups, resulting in an average accuracy increase of 9.7% compared to unimodal features. Furthermore, studies have identified statistically significant correlations between specific behavioural patterns and mental health outcomes, with two studies, for example, reporting consistent associations between mobility patterns and symptoms of anxiety and depression^{19–21}. However, the practical and scientific value of these intermediate outcomes may depend on their validation by clinical experts^{30,31}—thereby advancing the field of clinical psychology and neuropsychiatry of anxiety and depression—or on their translation into actionable insights that support user-facing applications.

In this article, we aim to address the following research questions: (i) How are variations in routines across different life aspects associated with self-reported anxiety and depression? (ii) Can personalised modelling based on routine variability identify specific routines that correlate with anxiety and depression? (iii) Does population-level grouping based on similar routine variability patterns reveal statistically significant between-group differences in mental health measures? (iv) Can an LLM translate modelling results into more accessible language to interpret an individual's routine variability pattern linked to anxiety and depression?

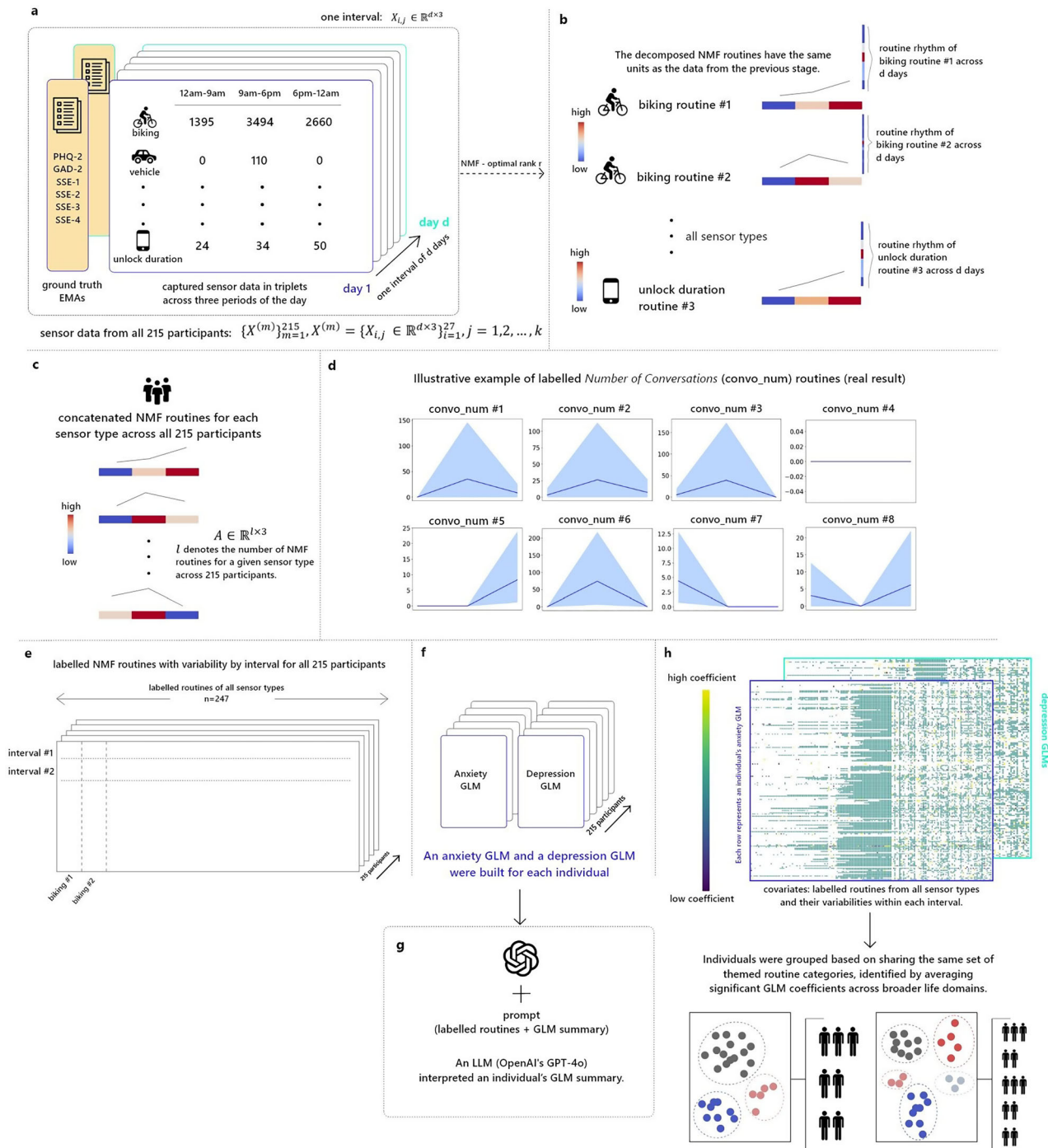
We conducted experiments on the recently released dataset College Experience Study³². This is the most extensive longitudinal mobile study to date, with automated sensing and survey data on 215 undergraduate participants at Dartmouth College across four years (2017–2020). Ecological momentary assessments (EMA) were prompted approximately weekly, including Patient Health Questionnaire 4 (PHQ-4)³³ and State Self-Esteem Scale (SSE)³⁴. As PHQ-4 includes PHQ-2 for depression and GAD-2 (Generalized Anxiety Disorder 2-item) for anxiety, we will refer to the items in PHQ-4 as PHQ-2 and GAD-2 for the rest of this article. Key findings from 4 associated studies experimenting with this dataset include: (1) a study³⁵ utilised mobile sensing and surveys to reveal that first-generation students exhibited distinct behavioural patterns, including reduced social interaction and higher academic engagement, correlating with elevated stress levels compared to peers; (2) a study³² compared pre-pandemic and pandemic-era behaviours, identifying reduced physical mobility, increased screen time, and declines in academic performance during COVID-19 lockdowns through GPS and app-usage data; (3) a third study³⁶ integrated fMRI data with mobile sensing, demonstrating that variations in students' daily routines (e.g., sleep irregularity and social isolation) predicted altered functional connectivity in brain networks associated with stress regulation; (4) a further study³⁷ analysed EMA responses and sensor data during COVID-19 initial phase, linking prolonged indoor mobility and disrupted sleep patterns to heightened anxiety and depressive symptoms. We believe that exploring the associations between routine variability and self-reported anxiety and depression presents a useful opportunity. A primary goal in digital twin (DT) and digital phenotype (DP) research in wellness is to develop a full-looped artefact to improve user well-being, bridged with human-computer interaction (HCI) research¹⁵. This intermediate-level knowledge could inform the design of digital artefacts focused on clarifying the locus of control (LoC)³⁸ by providing users with insights into how their daily behaviours affect their well-being. LoC reflects whether a person perceives a causal relationship between their own behaviour and rewards. Individuals with an internal LoC tend to attribute the outcomes of events to their own actions, which is shown to positively correlate with better life satisfaction and subjective well-being³⁸. In contrast, individuals with an external LoC tend to attribute outcomes to uncontrollable factors. It is challenging for individuals to gain insights into how their daily actions or behaviours are associated with well-being without monitoring behavioural data. Our approach and results inform future designs of DT agents for wellbeing that help users understand the controllable factors (e.g., exercise routines, phone usage) and uncontrollable factors (e.g., commuting to work) that contribute to their health.

An individual might exhibit a set of distinctive yet dominant routines for a given life aspect, such as exercising or socialising. For a student or an employee, a week of behavioural data might reveal two dominant routine types—those on weekdays and those on weekends—while the intensity of each routine type varies throughout the week. Computationally, the behavioural data over a period could be approximated by a weighted sum of several common routine types using dimension reduction techniques. Early work on uncovering such behavioural structures applied principal component analysis (PCA) to obtain a set of eigen behaviours and successfully inferred community affiliations with 96% accuracy based on individual and group-level behavioural similarities⁹. However, PCA forces the reduced components to be orthogonal to each other, assuming independence or uncorrelatedness between each routine, which does not align with actual human behaviour in the wild. Alternatively, non-negative matrix factorisation (NMF) learns a more interpretable part-based representation of the data while constraining the decomposed factors to be non-negative. For instance, NMF has been used in oncology to uncover subtypes in gene expression profiles^{39,40} and multi-omics data⁴¹.

We initiated our pipeline (see Fig. 1) by applying NMF on an interval's behavioural data between two consecutive EMAs prompts (approximately weekly). For each sensor type, we obtained a list of decomposed routines and corresponding day-to-day intensity over the interval. To answer research question (i), we conducted two-sample t-tests to identify significant differences in the standard deviation of routine intensity across the interval between low and high anxiety, and between low and high depression. We found statistically significant correlations in several data categories. As routines need to be relabelled in order to reach compromises to perform population-level comparison, we labelled routines by grouping them based on shape per sensing category using population-level quantile binning of routine intensity triplets, and then constructed a final matrix mapping each routine's variability to its corresponding label across all significant routine types. To answer research questions (ii) and (iii), two generalised linear models (GLMs) were built for each individual to associate routine variability of various data categories and anxiety or depression. Individuals were grouped by shared high-impact data categories.

After NMF decomposes the behavioural data into specific routine behaviours and their intensity, an individual profile on the correlation between the variability of specific routine behaviours and anxiety or depression could be obtained. For instance, two individuals with similar occupations might exhibit similar typical conversation routines, but how the variability of such routines over approximately a week affects their mental health might differ. Regression models, such as generalised linear models (GLMs), could be used to assess the individual-specific associations between routine variability and anxiety or depression. ANOVA was subsequently used to reveal between-group differences in mental health measures (PHQ-2, GAD-2 and each of the four items of SSE). Our result might inform group-based interventions. A previous study using eigen behaviour successfully inferred community affiliations within the population social network with 96% accuracy⁹. Another study that used smartphone data to monitor suicidal thoughts identified five phenotypes with latent profiles⁴². Furthermore, endophenotypes⁴³ might emerge, which are measurable characteristics representing underlying biological processes of a disease. However, further analyses are required to determine whether these subgroups satisfy the criteria for novel endophenotypes, including associations with specific gene regions, heritability, and disease association parameters⁴⁴.

While individual GLMs reveal certain patterns linking routine variability to anxiety or depression, we employed an LLM (GPT-4o (temperature=0, max token = 4096)) to translate the GLM results into more accessible language that could uncover actionable insights to facilitate self-regulation. There is increasing work in using LLMs in personal health domains; many are examining their predictive capabilities. For example, a study from Google⁴⁵ is a pioneering work that built a Personal Health Large Language Model (PH-LLM), fine-tuning the Gemini model to reason over time-series behavioural data for analysis and recommendations in sleep and fitness. Another study⁴⁶



evaluated multiple LLMs on various mental health predictive tasks and found promising yet limited performances.

Apart from understanding LLM's predictive capabilities for personal mental health, given the conversational nature of LLMs, we might be in the early stages of seeing LLMs as digital twin agents to assist self-regulation. A CHI study⁴⁷ examining the perceived benefits of digital mental health interventions (DMHIs) revealed that participants seeking active symptom management envisioned these tools functioning akin to a trusted companion—providing symptom awareness and fostering accountability. Additionally, a study by MIT Media Lab⁴⁸ employed AI as an interactive digital twin system where users converse with an AI-generated future self and reported reduced anxiety and increased motivation. Furthermore, another study at the

MIT Media Lab employed an LLM to enable users to query correlations, trends and actionable recommendations derived from Fitbit-captured physiological data. Participants in the user study reported that the model's outputs were informative and practically useful⁴⁹. In our study, the LLM interpretation of individual GLMs also shows a potential for users to engage with their own routine-based digital phenotypes. While reviewing how specific routines correlate with their moods, users might self-reflect and consider regulating controllable behaviours (as opposed to uncontrollable factors, for instance, when they learn that a long-time morning commute to school correlates with their mood). We see future research direction of framing LLMs as a self-reflective, data-driven digital twin agent to make sense of personal behavioural data and facilitate self-regulation.

Fig. 1 | The diagram illustrating the analysis pipeline. **a** The initial dataset consists of behavioural data from 215 participants across 27 mobile sensing categories in time intervals; an interval is defined as the period between two consecutive EMAs. The end-of-interval EMAs were taken as the ground truth mental health measures for the interval. All sensor data were originally in the format of three periods of the day (12 am–9 am, 9 am–6 pm, 6 pm–12 am) and were normalised (e.g., hour and second units transformed into minutes). **b** NMF was applied separately to each sensor data type for each interval, decomposing the data into a set of routine behaviours and their corresponding intensities over the interval. The number of routines was determined through leave-one-out cross-validation (LOOCV), yielding the lowest reconstruction error. The NMF routines retain the same data units as the original data. Routine variability was computed as the standard deviation of routine intensity across the interval. **c** For each sensing category, routine behaviours from all 215 participants were concatenated into a matrix to identify similar patterns for labelling. Specifically, the 25%, 50%, and 75% quantiles of each triplet item (corresponding to each period of the day) were computed, and labelling was based on the unique

combinations of these results. This means a given sensing category could have a maximum of 27 routine types (3^3). **d** An example of labelled routines for Number of Conversations (convo_num). Routine #1 means the participant had zero conversations in the morning (before 9a m), fewer than 50 (but more than zero) during the day (9 am–6 pm), and fewer than 25 in the evening (after 6 pm). **e** New matrices were obtained for all 215 participants' labelled NMF routines for each interval. **f** For each participant, a generalised linear model (GLM) for anxiety and another for depression were built to associate routines and their variability with PHQ-2 and GAD-2 scores, respectively. **g** A large language model (OpenAI's GPT-4o) interpreted an individual's GLM summary to generate practical insights. **h** Subsequently, individuals were grouped if their GLMs shared the same set of themed routine categories, identified by averaging significant GLM coefficients across broader life aspects. Certain parts of this figure come from sources include Wikimedia Commons (the logo of OpenAI), The Noun Project (<https://thenounproject.com>) (icons indicating participants and sensing categories) under the Creative Commons Attribution 3.0 license.

Results

The rank of the NMF decomposition was constrained to be no greater than the minimum of 3 (indicating the original data format of three daily periods) and d (the number of days in the interval). Since the interval spans approximately a week ($d \approx 7$) and is typically longer than 3 days, the possible ranks were limited to 1, 2, or 3. Most NMF decompositions resulted in a rank of 1 (p (rank = 1) = 99.76%, p (rank = 2) = 0.2%, p (rank = 3) = 0.04%), indicating that for most participants, a single dominant routine behaviour was sufficient to reconstruct approximately a week's worth of behavioural data.

Routine variability captured from certain sensor categories correlated with the presence of high anxiety or high depression

After obtaining decomposed NMF routines and their variability within each interval for all 27 sensing categories from 215 participants, two-sample t -tests were applied to each combination of sensing category and PHQ-2 or GAD-2 scores to examine population-level correlations between routine variability and the presence of high anxiety or high depression. The t -tests revealed that routine variability in the *number of conversations*, *audio voice*, and *conversation duration* was negatively correlated with both anxiety and depression (see Table 1). This suggests that individuals with more stable routines in these life aspects were more likely to experience symptoms of anxiety or depression. In contrast, routine variability in the *duration of phone-played audio sessions*, *duration of phone usage when unlocked*, and *number of phone locks/unlocks* showed positive correlations with both anxiety and depression, indicating that greater variability in these behaviours was associated with higher symptom levels.

Additionally, individuals were more likely to experience anxiety when routines related to biking and duration of being still were more stable, and when routines involving the number of incoming SMS, standard deviation and mean of audio amplitude, and walking were more variable. For depression, significant negative correlations were observed with variability in routines related to running, duration of incoming and outgoing calls, number of outgoing calls, and number of location visits.

An LLM translates an individual's anxiety GLM results into user-accessible language to uncover how routine variability is associated with mental health measures and personalized actionable insights

Figure 2 presents the LLM output interpreting an individual's anxiety GLM summary. The LLM was able to extract statistically significant patterns and describe the effects of variability based on routine definitions. It demonstrated a potential to make actionable suggestions to support self-regulation for individuals based on the routine variability pattern grounded in evidence-based open-source consume health information.

Grouping individuals based on similar routine variability correlation revealed significant between-group differences in certain mental health measures

All data categories ($n = 27$) in which routine variability showed statistically significant correlations with PHQ-2 or GAD-2 were included for labelling (12 categories for anxiety and 13 for depression), enabling comparison of NMF-derived routines across individuals. Routines from the same sensing category were labelled based on their shape. For example, two biking routines showing biking activity in the morning and no activity during the rest of the day would receive the same label. To achieve this, we computed the 25% (low), 50% (mid), and 75% (high) quantiles of population-level routine values and assigned labels based on their unique combinations. For instance, (low, mid, low) and (low, high, mid) represent two distinct routine behaviours and were assigned different labels. Each sensing category therefore, yielded a maximum of 27 unique labels. On average, each sensing category for anxiety produced 20.58 unique labels (247 in total), and each category for depression produced 17.47 labels (227 in total). After labelling all routines from all participants, we obtained a final-stage pre-processed matrix for each individual in which the NMF-derived routines for each interval were labelled. Each row represents one interval, with the original behavioural data decomposed and labelled into specific routine types and their corresponding variability across the various data categories. We then built a generalised linear model (GLM) with a binomial distribution for each individual, serving as a personalised model for both anxiety and depression.

For each individual, we obtained a list of routines where their variability significantly correlated with PHQ-2 or GAD-2 scores. For instance, we discovered that the variability of routine bike#3 and routine unlock_duration#2 are associated with an individual's GAD-2 scores over approximately four years of data. As we are interested in understanding if there are commonalities in individuals if they share the same routine variability correlation patterns with PHQ-2 or GAD-2, we proceeded to group individuals based on their GLM results. However, due to the large number of unique routine types, the initial grouping did not identify any groups with more than one participant. Therefore, for this analysis, we chose to consider only the broad categories, as well as removing redundancy for those that reflect similar life aspects (e.g., conversation duration and number of conversations), resulting in larger group membership.

Notably, GLMs are deemed unsuitable for individuals who did not have any high anxiety or depression, as the response variable contained only zeros. There were 35 such individuals for anxiety models (with GAD-2 score as the response variables) and 50 for depression models (with PHQ-2 score as the response variables). Furthermore, 140 individuals for anxiety models and 128 for depression models did not have any significant routines and were thus grouped together. Ultimately, a total of 6 groups emerged for anxiety models, including the 2 groups with non-high GAD-2 scores and with no significant GLM results. For depression models, a total of 4 groups

Table 1 | Two-sample t-tests for correlations between population-level routine variability by sensing category and anxiety (low or no anxiety, high anxiety) & depression (low or no depression, high depression)

Category ^a	df ^b	Anxiety t-statistics (CI)	P value	Depression t-statistics (CI)	P value
vehicle	32,341	0.32 (−0.62, 0.87)	0.7497	−0.61 (−0.99, 0.52)	0.5429
bike	32,341	−2.58 (−1.0, −0.14)	0.0098**	−1.77 (−0.83, 0.04)	0.0767
foot	6575	0.06 (−1.15, 1.23)	0.9485	0.39 (−0.83, 1.25)	0.6956
running	32,341	0.94 (−0.14, 0.39)	0.3498	−3.1 (−0.69, −0.16)	0.0019**
still	32,341	−2.04 (−0.75, −0.02)	0.0413*	0.01 (−0.37, 0.38)	0.994
walking	32,341	6.43 (1.19, 2.24)	<0.001***	1.43 (−0.14, 0.92)	0.1525
audio_amp_mean	10,841	3.39 (1.84, 6.85)	<0.001***	−1.38 (−4.14, 0.72)	0.1678
audio_amp_std	10,841	3.34 (1.48, 5.69)	<0.001***	−0.66 (−2.72, 1.35)	0.5091
convo_duration	32,341	−6.63 (−2.89, −1.57)	<0.001***	−3.6 (−1.9, −0.56)	<0.001***
convo_num	32,341	−8.31 (−0.1, −0.06)	<0.001***	−3.9 (−0.06, −0.02)	<0.001***
audio_voice	32,341	−8.07 (−0.02, −0.01)	<0.001***	−4.9 (−0.01, −0.01)	<0.001***
call_in_duration	6575	0.43 (−0.67, 1.05)	0.6642	−2.81 (−1.82, −0.32)	0.0050**
call_in_num	6575	−0.23 (−0.05, 0.04)	0.8166	−1.44 (−0.07, 0.01)	0.1495
call_miss_num	6575	0.54 (−0.01, 0.01)	0.5924	−1.11 (−0.01, 0.0)	0.2662
call_out_duration	6575	−0.92 (−1.28, 0.46)	0.3559	−2.66 (−1.8, −0.27)	0.0079**
call_out_num	6575	−0.6 (−0.06, 0.03)	0.5458	−2.43 (−0.1, −0.01)	0.0152*
light_mean	5869	−1.09 (−0.66, 0.19)	0.2763	0.73 (−0.24, 0.53)	0.4632
light_std	5869	0.6 (−48.33, 91.09)	0.5477	0.29 (−53.77, 72.8)	0.7682
loc_dist	27,339	−0.32 (−20.79, 14.92)	0.7474	−1.5 (−31.97, 4.21)	0.1325
loc_max_dist_from_campus	27,325	−0.41 (−20.23, 13.24)	0.6826	−1.5 (−29.94, 3.96)	0.1331
loc_visit_num	274,23	−1.66 (−0.02, 0.0)	0.0976	−3.31 (−0.02, −0.01)	<0.001***
other_playing_duration	25,775	2.92 (0.5, 2.56)	0.0035**	4.86 (1.58, 3.72)	<0.001***
other_playing_num	25,775	1.2 (−0.0, 0.02)	0.229	2.41 (0.0, 0.03)	0.0158*
sms_in_num	6575	2.27 (0.02, 0.25)	0.0234*	−0.12 (−0.11, 0.09)	0.9075
sms_out_num	6575	1.14 (−0.05, 0.18)	0.2547	−3.89 (−0.29, −0.1)	<0.001***
unlock_duration	32,341	2.92 (0.21, 1.08)	0.0035**	9.14 (1.61, 2.49)	<0.001***
unlock_num	32,341	2.66 (0.01, 0.07)	0.0077**	5.47 (0.06, 0.12)	<0.001***

^aThe full descriptions and supported operation system (OS) of mobile sensing categories are presented in Table 5.

^bDegrees of freedom.

P-value < 0.05*; P value < 0.01**; P value < 0.001***.

emerged, including the ones with all non-high PHQ-2 scores and non-significant GLM results.

In Table 2, we show the pairwise comparisons of anxiety models with Tukey's test for a total of 5 groups, excluding the one group in which individuals are known to contain non-high GAD-2 scores and is hence deemed to be unsuitable to compare with other groups on mental health measures. For the 4 groups with significant GLM results, there are 4 mutual sensing categories: *standard deviation of audio amplitude*, *biking*, *duration of phone-played audio sessions* and *duration being still*. 11 participants were present in group #1 which additionally contained *conversation duration*, *number of incoming SMS*, and *walking*. Group #2 is individuals with non-significant GLM results. 20 participants were present in group #3, which additionally contained *conversation duration* and *walking*. 7 participants were present in group #4, which additionally contained *conversation duration* and *the number of incoming SMS*. 2 participants were present in group #5 which additionally contained *audio voice* and *walking*. One-way ANOVA revealed significant differences in GAD-2 scores ($p = 0.0078$); no significant results were found for PHQ-2 scores and any items of SSE (see Table 3). Prior to applying ANOVA, we conducted Levene's test for homogeneity of variance and Shapiro-Wilk tests for normality to justify the use of ANOVA (all null hypotheses were rejected due to insignificant p-values). The same assumption testing was also carried out for depression

models. Post-hoc Tukey's test identified that the difference lay between group #2 and group #3.

For depression models with significant GLM results, 2 groups (group #2 with 6 individuals and group #3 with 31 individuals) emerged, both containing the *number of incoming calls*, *duration of outgoing calls*, *conversation duration*, *number of local visits*, and *other playing duration*. Group #1 are individuals with no significant GLM results. In Table 4, we showed the pairwise comparison results with Tukey's test for all 3 groups, further including group #1. Group #2 additionally contained the *number of outgoing SMS*. One-way ANOVA revealed strong significant between-group differences in GAD-2 scores ($p < 0.001$), PHQ-2 scores ($p < 0.001$), SSE-2 (satisfaction with appearance) ($p = 0.0176$), SSE-3 (feeling as smart as others) ($p < 0.001$), and SSE-4 (overall feeling about oneself) ($p < 0.001$). All post-hoc Tukey's tests revealed that the difference lay between group #1 and group #3.

Discussion

To our knowledge, little research exists that explores the links between routine behaviours and routine variability with anxiety and depression while considering multiple life aspects using longitudinal behavioural data. Our findings clarify these links by distilling complex behavioural data into individual-specific routines. We demonstrated that captured

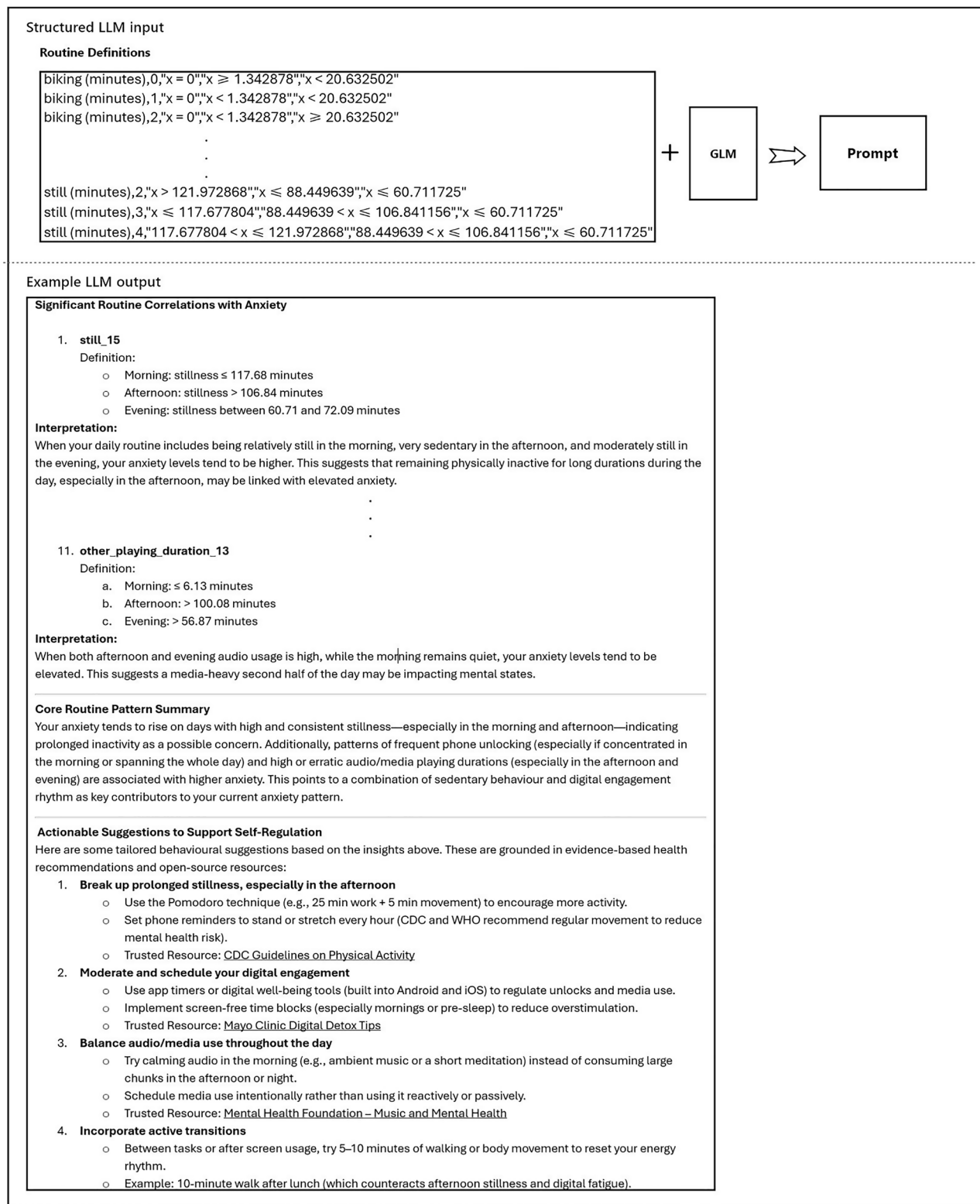


Fig. 2 | Interpretation of Individual GLM Results Using an LLM (GPT-4o). We used an LLM (OpenAI's GPT-4o (temperature=0, max token=4096)) to translate an individual's anxiety GLM results into user-accessible language along with the routine definition file. We presented the actual output from the LLM.

mobile sensing data could be decomposed into a set of individual-specific routines for three periods of the day and their weekly variability using NMF. Importantly, routine variability is computed as the standard deviation in routine weights and should be distinguished from routine regularity. A regularly changing routine with a small amplitude will have much smaller routine variability compared to one with a large

amplitude. As NMF results in a partial representation of the factors, our approach emphasises the change in routine behaviours on different days, penalising cases of regular but similar routines. For instance, an individual with drastically different weekday and weekend routines will exhibit greater variability compared to someone with similar, albeit regular, weekday and weekend routines.

We found population-level statistically significant associations between routine variability and mental health outcomes (anxiety and depression symptoms, as reported in EMAs), though the direction of associations depends on the sensing category. Conversation and audio voice-related routines had the strongest negative correlations with both anxiety and depression, while routines related to phone usage showed positive correlations with both anxiety and depression. Additionally, mobility-related routines exhibited more variability when associated with anxiety, but less variability when associated with depression. Routines related to outgoing SMS and call durations were more stable when associated with depression. Conversation routines, detected through sensing categories that capture ambient voice, reflect an individual's real-world social life and should be considered separately from digital communications, such as texting, calling, or using social media. Furthermore, ambient audio amplitudes are likely to change when a person moves between places on foot, so relevant routines might reflect similar life aspects related to mobility and surrounding environments. Overall, the results revealed that both anxiety and depression are associated with more stable patterns of real-

world socialisation but more variable phone usage patterns, reflected by duration of usage, locks and unlocks and phone-played audio sessions. Anxiety is associated with more variable mobility-related routines, whereas depression is associated with more stable routines related to mobility and calls.

Current research has focused more on social rhythmicity (regularity) and bipolar disorders^{50–52}, with less emphasis on the links with anxiety and depression⁵. Most relevant research has found that greater regularity is associated with more positive mental health outcomes, often relying on Likert-scale questionnaires for participants to reflect on past behaviours⁵. Our findings suggest a more nuanced view of the links between social routines and both anxiety and depression, considering real-world socialisation and digital communications (calling and texting) separately, based on analysis of years-long continuous behavioural data instead of questionnaires.

A recent paper published in 2024 highlighted that little is known about the impacts of anxiety on people's travel². Other studies have also suggested that mental health problems play a role in people's transportation choices^{53,54}. Our results reveal that anxiety tends to result in more variable on-foot mobility patterns, meaning that how long anxious individuals spent walking at different times of the day varied greatly from day to day. In contrast, depressed individuals tended to have more stable mobility patterns in terms of the number of location visits and the duration of running. This finding aligns with several existing digital phenotyping studies. Furthermore, some previous studies have identified that longer home stays reflect depressive symptom severity^{3,55}, and some has found that physical exercise has antidepressant effects⁵⁵. Moreover, one known symptom of depression is the loss of interest in enjoying activities⁵⁶. The negative correlation between depression and variability in mobility routines indicates that more depressed individuals move at a similar time across days. A potential interpretation is that when a student consistently visited places that might have fixed schedules (e.g., attending lectures and having meals) to maintain minimum needs, such behaviours could lead to an increase in depressive symptoms. Notably, all significant data categories related to the acts of walking or running; no significance was found with vehicle-taking routines. We hypothesise that taking vehicles is often a passive act affected by uncontrollable circumstances (e.g., commuting between home and school or work) and might reflect less about a person's mental well-being.

For each participant, a GLM model was built to reveal whether there were any significant routines, with the presence of anxiety or depression as the response variables. Individuals were then grouped based on broad categories, and those that reflected similar life aspects (e.g., number of conversations, duration of conversations) were removed. Grouping individual anxiety models, ANOVA found significant between-group differences in GAD-2 scores but not in any other mental health outcomes. Grouping individual depression models, strong between-group differences were found in anxiety, depression, SSE question 2 (satisfaction with

Table 2 | Tukey's test for pairwise mean comparisons between groups for anxiety models

Anxiety			
Group A	Group B	MeanDiff (CI)	P-adj
1	2	−0.57 (−1.31, 0.17)	0.2184
1	3	0.08 (−0.81, 0.97)	0.9991
1	4	−0.69 (−1.83, 0.46)	0.4644
1	5	−0.03 (−1.85, 1.79)	1
2	3	0.65 (0.08, 1.21)	0.0156*
2	4	−0.12 (−1.04, 0.8)	0.9965
2	5	0.54 (−1.15, 2.22)	0.904
3	4	−0.77 (−1.81, 0.27)	0.253
3	5	−0.11 (−1.87, 1.65)	0.9998
4	5	0.66 (−1.24, 2.56)	0.8752

Group 1 contains significant routines of: standard deviation of audio amplitude, biking, conversation duration, duration of phone-played audio sessions, number of incoming SMS, duration of being still, walking.

Group 2 contains individuals with no significant routines found.

Group 3 contains significant routines of: standard deviation of audio amplitude, biking, conversation duration, duration of phone-played audio sessions, duration of being still, walking.

Group 4 contains significant routines of: standard deviation of audio amplitude, biking, conversation duration, duration of phone-played audio sessions, duration of being still, number of incoming SMS.

Group 5 contains significant routines of: standard deviation of audio amplitude, audio voice, biking, duration of phone-played audio sessions, duration being still, walking.

P value < 0.05*.

Table 3 | One-way ANOVA results of between-group differences in mental health measures (PHQ-4 scores and SSE-scores) by grouping individual GLM results

EMA ^a	Anxiety			Depression		
	df ^b	f-statistics	P value	df ^b	f-statistics	P value
GAD-2	4,175	3.58	0.0078**	2,162	10.38	<0.001***
PHQ-2		1.46	0.215		15.94	<0.001***
SSE-1		1.76	0.138		2.64	0.0744
SSE-2		0.33	0.858		4.14	0.0176*
SSE-3		1.11	0.352		8.96	<0.001***
SSE-4		0.92	0.455		7.73	<0.001***

^aThe full description of EMAs are presented in Table 6.

^bDegrees of freedom.

P value < 0.05*; P value < 0.01**; P value < 0.001***.

Table 4 | Tukey's test for pairwise mean comparisons between groups for depression models

Group A	Group B	PHQ-2		GAD-2		SSE-2		SSE-3		SSE-4	
		MeanDiff (CI)	P-adj	MeanDiff (CI)	P-adj	MeanDiff (CI)	P-adj	MeanDiff (CI)	P-adj	MeanDiff (CI)	P-adj
1	2	0.75 (−0.09, 1.58)	0.0904	0.08 (−0.77, 0.92)	0.9743	−0.33 (−0.98, 0.32)	0.4507	−0.43 (−1.05, 0.18)	0.2206	−0.48 (−1.07, 0.12)	0.1423
1	3	0.92 (0.52, 1.32)	<0.001***	0.78 (0.37, 1.18)	<0.001***	−0.36 (−0.67, −0.46)	0.0201*	−0.5 (−0.8, −0.2)	<0.001***	−0.44 (−0.72, −0.15)	0.0012**
2	3	0.17 (−0.72, 1.06)	0.8947	0.7 (−0.2, 1.6)	0.1603	−0.03 (−0.72, 0.67)	0.9958	−0.07 (−0.73, 0.59)	0.9673	0.04 (−0.59, 0.68)	0.9673

Group 2 contains significant routines of: duration of incoming calls, duration of outgoing calls, conversation duration, number of location visits, duration of phone-played audio sessions, and number of outgoing SMS.

Group 3 contains significant routines of: duration of incoming calls, duration of outgoing calls, conversation duration, number of location visits, and duration of phone-played audio sessions.

No significance in SSE-1 scores.
P value < 0.05*; P value < 0.01**; P value < 0.001***.

appearance), SSE question 3 (feeling as smart as others), and SSE question 4 (overall feelings about oneself), but not in SSE question 1 (worrying about what others think of them). More than half of all participants did not have any specific routine where its variability was significant to anxiety or depression over time. Around a third of the population was found to have variabilities in certain routines correlated with anxiety or depression in the long run. Post-hoc ANOVA revealed that for anxiety models, the difference lay between the group with no significant routines and the group with significant routines related to audio amplitude, biking, conversation duration, phone-played audio sessions, duration being stationary, and walking. Post-hoc ANOVA for depression models revealed that the difference in all significant mental health outcomes lay between the group with no significant routines and the group with significant routines related to the number of incoming calls, duration of outgoing calls, conversation duration, number of local visits, and other playing duration.

Research in DT for wellbeing is still in its early stages¹⁵. When developing DTs to target the concept of wellbeing, which holistically encompasses both physical and mental health, it is necessary to bridge knowledge from several disciplines, primarily clinical psychology, behavioural science, human-computer interaction (HCI), and machine learning (ML) or artificial intelligence (AI). A DP or DT agent that uncovers an individual's specific routines and clarifies how these routines interact with their mental health states in terms of intensity, regularity, and variability, has the potential to internalise or reorient one's locus of control (LoC)³⁷. Our study showed the technical plausibility of using an LLM to translate an individual's GLM result into specific actionable insights. The role of LLM servers as an interpreter rather than freely generate content to minimise the known issue of hallucination, hence it could translate machine learning or data mining-derived insights about a user's behavioural data – the pattern of routine variability is one example. We can imagine a more complete LLM-based DT agent for healthy behavioural self-regulation that includes components on (1) explaining behavioural data, (2) predicting mental health states (3) explaining predictions and deriving actionable insights (4) proactively notifying users for healthy behavioural change based on an individual's historical data. When specific patterns about how a user's daily behaviours associate with their mental health outcomes are uncovered, it first raises their self-awareness and might start self-regulating the controllable ones. It is up to the user to decide which behaviour has the potential to be adjusted and which is beyond their control. Further user studies with HCI principles are needed to evaluate the perceived usefulness of outputs from LLM-based DT agents, as well as to evaluate performance of different language models.

The grouping analysis also potentially informs future work on group-based treatments, similar to user-based collaborative filtering for recommendation systems³⁵. Specifically, recommendations could be made to users based on their effectiveness for others with the same group membership.

Several limitations should be noted. First, the dataset includes students from a single U.S. college, limiting generalisability to other age groups, socio-economic strata, or cultural contexts. Second, the data collection spanned the COVID-19 pandemic, which introduced atypical behavioural patterns; findings may not be generalised to non-pandemic conditions. Third, many NMF decompositions resulted in rank-1 routines, which might seem to contradict the notion of multi-routine; we hypothesize that once extending the duration of intervals NMF will decompose into more routines (e.g., examining bi-weekly or monthly routine behaviours). Fourth, GLMs could not be fitted for a considerable portion of participants due to zero-variance outcome variables, excluding them from subgroup analysis. Fifth, the quantile-based binning strategy used for routine labelling is sensitive to this specific cohort and may not generalise to other datasets. Sixth, the initial grouping based on fine-grained routine types did not yield groups larger than one participant. This is likely due to the number of participants ($n = 215$) being still relatively small compared to labelled routines ($n > 200$). For this study, reducing grouping by broad categories enabled between-group analysis, limiting the resolution of findings. The grouping of participants by broad behavioural categories may obscure important within-category routine differences. Seventh, the ethical implications of long-term behavioural

monitoring warrant critical attention. Although passive sensing offers rich behavioural insights, prolonged tracking of individuals may raise privacy, autonomy, and informed consent concerns. Eighth, a total of 54 independent two-sample t-tests were conducted. Given the exploratory nature of this analysis, corrections for multiple comparisons were not applied. Therefore, the reported p-values should be interpreted with caution, and the findings considered as hypothesis-generating rather than confirmatory. Future work should integrate ethical guidelines for the design and deployment of such systems, and include user-centred studies to ensure transparency, perceived control, and consent are respected. Lastly, we used the LLM to offer a demonstration that the modelling results could be expressed in more user-accessible language to potentially facilitate self-regulation and behavioural awareness. How users would perceive and engage with their routine-based digital phenotypes remains a topic for future research.

In conclusion, our experiment contributes to the field of DT or DP for mental health by exploring the behaviours associated with depression and anxiety in the wild in terms of specific routine behaviours and their variability. We further demonstrate the technical plausibility of an LLM to translate the correlation analysis into more user-accessible language for specific-actionable insights to facilitate self-regulation. We successfully used NMF to decompose the data into routines and their weights and found significant correlations between mental health outcomes and routines primarily related to social life and mobility. We pointed out our further research directions in developing DTs to target well-being. We discussed a DT/DP agent that clarifies the

controllable or uncontrollable factors contributing to health has great potential to strengthen one’s internal locus of control.

Methods

College Experience Study Dataset

The recently released dataset College Experience Study³² is the most prolonged mobile study to date, with automated sensing and survey data on 215 undergraduate participants at Dartmouth College across four years (2017–2020). The dataset is released in stages and the raw sensing data has not been released by the time this manuscript is submitted. Hence, we conducted experiments on the most recently released daily and hourly aggregated sensing data and ecological momentary assessments (EMAs) in March 2024. The dataset is publicly accessible on Kaggle. Table 5 presents the data dictionary of all 27 mobile data sensing categories from the College Experience Study dataset that are in the format of 3 periods of the day (12 am–9 am; 9 am–6 pm; 6 pm–12 am). Descriptions and the supported mobile phone operating system (OS) of each category are also presented. Table 6 presents the two primary EMAs (PHQ-2, GAD-2, and SSE) included in the dataset.

Decomposing weekly behavioural data into common routines with non-negative matrix factorisation

The routine analysis was conducted on all data categories ($n = 27$), each in format of a numeric triplet representing 3 periods of the day. We chose the

Table 5 | Full descriptions and supported operating systems of each mobile sensing category of a total of 27

Category	Full Description	Operation System (OS)
vehicle	Activity duration in vehicle	All
bike	Activity duration on bike	All
foot	Activity duration on foot	Android only
running	Activity duration running	iOS only
still	Activity duration still (no movement)	All
walking	Activity duration walking	iOS only
audio_amp_mean	Mean audio amplitude	Android only
audio_amp_std	Standard deviation of audio amplitude	Android only
convo_duration	Duration of detected conversations	Android only
convo_num	Number of detected conversations	Android only
audio_voice	Ratio of detection of voice to number of inferences made during the period. Indicates, of all the inferences made during this specific period, the percentage with voice present (as opposed to noise).	Android only
call_in_duration	Duration of incoming call	Android only
call_in_num	Number of incoming calls	Android only
call_miss_num	Number of missed calls	Android only
call_out_duration	Duration of outgoing calls	Android only
call_out_num	Number of outgoing calls	Android only
light_mean	Light mean amplitude	Android only
light_std	Light amplitude standard deviation	Android only
loc_dist	Distance travelled	All
loc_max_dist_from_campus	Maximum distance travelled away from campus	All
loc_visit_num	Number of locations visited	All
other_playing_duration	This metric tracks the total length of audio sessions, including any form of audio-based media like music or video, being played on the phone. It tallies the cumulative time spent on audio playback.	iOS only
other_playing_num	This metric tracks the frequency of audio sessions, including any form of audio-based media like music or video, being played on the phone. It tallies the number of times audio was played.	iOS only
sms_in_num	Number of incoming SMS	Android only
sms_out_num	Number of outgoing SMS	Android only
unlock_duration	Average amount of time that a user uses the phone in unlocked state (i.e., total phone usage duration)	All
unlock_num	Number of lock/unlocks	All

Table 6 | full descriptions of EMA questions

EMA	Description
PHQ4-1	Over the last 2 weeks, how often have you been bothered by the following problems? Feeling nervous, anxious or on edge
PHQ4-2	Over the last 2 weeks, how often have you been bothered by the following problems? Not being able to stop or control worrying
PHQ4-3	Over the last 2 weeks, how often have you been bothered by the following problems? Feeling down, depressed or hopeless;
PHQ4-4	Over the last 2 weeks, how often have you been bothered by the following problems? Little interest or pleasure in doing things
SSE-1	Right now, I worry about what other people think of me.
SSE-2	Right now, I am pleased with my appearance.
SSE-3	Right now, I feel as smart as others.
SSE-4	Right now, Overall, I feel good about myself.

PHQ-4: Patient Health Questionnaire 4³³.

PHQ4-1 and PHQ4-2 are combined as GAD-2 for anxiety and PHQ4-3 and PHQ4-4 are combined as GAD-2 for depression.

SSE: State Self-Esteem Scale³⁴.

time window between two consecutive EMAs to decompose the behavioural data into a set of dominant routine(s). The data in original formats can be formalised as $\{X^{(m)}\}_{m=1}^{215}$, $X^{(m)} = \{X_{i,j} \in \mathbb{R}^{d \times 3}\}_{i=1}^{27}$, $j = 1, 2, \dots, k$, for a total of 215 participants across 27 mobile sensing categories and $d = T_{\text{EMA}_{i+1}} - T_{\text{EMA}_i}$ in days. For each interval, PHQ-4 and SSE scores on the end day $T_{\text{EMA}_{i+1}}$ were assigned to the interval.

EMAs used in this study contained PHQ-4 and SSE scores; they are used to define the mental health outcomes. As PHQ-4 combines GAD-2 (surveying anxiety) and PHQ-2 (surveying depression), we will use them to denote PHQ-4 scores for anxiety and depression respectively. GAD-2 scores of at least 3 is typically used to indicate a positive screen for generalised anxiety disorder. At this threshold, the GAD-2 has demonstrated a sensitivity of 86% and specificity of 83% for diagnosing generalised anxiety disorder⁵⁷. PHQ-2 scores of least 3 is considered optimal to identify individuals at risk for major depressive disorder. At this threshold, studies have reported a sensitivity of 74% and specificity 75% for detecting major depression⁵⁸. These two measures are used to conduct the correlation analysis and GLM as the target variables to answer research questions (i) and (ii). SSE, which are self-reported surveys of self-esteem, are only added in the analysis of examining between-group differences in mental health measures, which included all EMAs (GAD-2, PHQ-2 and four items of SSE).

For each interval, the behavioural data $X_{i,j} \in \mathbb{R}^{d \times 3}$ (from category i and from interval index j) is decomposed into a routine matrix $H_{i,j} \in \mathbb{R}^{3 \times r}$ and a routine intensity matrix $W_{i,j} \in \mathbb{R}^{d \times r}$ with non-negative matrix factorisation $\text{NMF} \approx WH^T$, $r \leq \min(d, 3)$, where r (rank) is the number of decomposed routine types within the time window for a given sensing category. For each individual NMF, the optimal rank is found with leave-one-out cross-validation (LOOCV), iterating from the range of 1 to $\min(d, 3)$. Specifically, in each iteration, NMF is trained on $d-1$ rows of X to reconstruct the hold-out day. The optimal rank is chosen as the one that yields the lowest mean reconstruction error (the Frobenius norm of the difference between the reconstructed matrix WH^T and X), averaging over d .

Correlation between routine instability and PHQ-4 scores

For each interval, the routine variability is computed as the standard deviation of the routine intensity array, $\text{col}(W_{i,j} \in \mathbb{R}^{d \times r})$, corresponding to the routine type, $\text{row}(H_{i,j} \in \mathbb{R}^{3 \times r})$. We are interested in understanding if there are statistically significant differences in means of routine variabilities of all routines between the low anxiety group and the anxiety group, and between the low depression group and the depression group. The low anxiety group contains all time windows with a GAD-2 score lower than 3, and the anxiety group contains all time windows with a GAD-2 score greater than or equal to 3. The same threshold of 3 is used to distinguish low

depression and depression groups. Two-sample t-test is individually conducted on each sensing category for anxiety and for depression, without regard to which individual each interval is from, resulting in 54 independent tests. Levene's test was used to assess equal variance; if the null hypothesis of equal variance was rejected, Welch's t-test, which does not assume population variance, was performed. We selected the subset of data categories that yielded significant p -values ($p < 0.05$) from t-test results for the next step in the analysis.

Population-wise routine labelling

For each data category, we concatenate the decomposed routines from all individuals into a population-level meta-routine matrix $A \in \mathbb{R}^{1 \times 3}$, with each row as a numeric triplet (x_0, x_1, x_2) indicating three periods during the day, and k is the total number of routines. Routines are subsequently labelled based on the population-level quantiles in which their values fall.

For each sensing category, let $Q_{0.25}$, $Q_{0.5}$, $Q_{0.75}$ be the 25th, 50th and 75th percentiles (quantiles) for A . We label each x of the routine with 'low' if $x_i \leq Q_{0.25}$, 'mid' if $Q_{0.25} < x_i \leq Q_{0.75}$ and 'high' if $x_i > Q_{0.75}$, given that $Q_{0.25}$, $Q_{0.5}$, $Q_{0.75}$ are all non-zeros. If $Q_{0.25}$ is zero and $Q_{0.5}$ is non-zero, we label each x of the routine with 'low' if $x_i \leq Q_{0.5}$ and 'high' if $x_i > Q_{0.5}$. If only $Q_{0.75}$ is non-zero, we label each x of the routine with 'low' if $x_i = 0$ and 'high' if otherwise. Subsequently, routines are grouped and labelled based on the unique combinations of low, mid and high. For each population-level routine type matrix A of any sensing category, the maximum number of routine labels is 27, as each value of the numeric triplet can be categorised as low, middle, or high. Importantly, the binning based on population quantiles is only used to label each routine based on shape; each individual routine from a time interval I is assigned with its variability. The final matrix has dimension $\mathbb{R}^{q \times v}$, where v is the total of time intervals from all participants, and q is the total number of dominant routine types ($q = 32368$; $v = 247$ for anxiety, $v = 227$ for depression) from all eligible data categories (of which previously the population-level routine variability is shown to be statistically significant with anxiety or depression) in addition to any EMA categories.

Individualised anxiety and depression models

We proceed to build personalised models to discover how the variability of each routine correlates with GAD-2 and PHQ-2 scores. Among all 27 sensing categories, 15 are Android-only, 4 are iOS-only and 8 are supported in both operating systems. We removed any excess columns where all values are missing. Considering the case that people might switch to from one OS to another over the years when having new phones, leaving some of the values in the columns to be empty, we filled in the missing values in such columns with zero. Subsequently, a generalised linear model (GLM) is applied to all data from each individual to assess the relationship between routine variability of labelled routines as predictor variables and the response variable, a Boolean indicating the presence of anxiety or depression. 35 individuals did not have any GAD-2 scores greater than or equal to 3, and 50 individuals did not have PHQ-2 scores greater than or equal to 3. GLMs model could not be built for those individuals. We retained a list of such individuals for later analysis.

LLM interpreting individual GLM summary

We employed an LLM (OpenAI's GPT-4o (temperature=0, max token=4096)) configured with a prompt. The prompt specifies the role (a helpful, empathetic agent that raises awareness of daily routine patterns), the task (translate statistically significant GLM associations between variability in labelled routines and anxiety measured by GAD-2 into clear, non-technical explanations and match each routine code to its definition file), the data inputs (pre-generated weekly insights, GLM outputs and a lookup file of routine definitions), the communication style (concise language focused on observable behaviours only), and a caution clause that mitigates off-topic or unsafe use. The model reports routine variability-anxiety associations without effect sizes or p -values. The LLM was prompted to emphasise real-world behaviours and encourages reflection by highlighting recurring

patterns across the day. It then synthesises the behavioural profile into a concise summary, identifying key anxiety-linked behaviours over the monitoring period. The recommendations were prompted by concrete actions (e.g., improving sleep consistency or reducing late-night screen exposure) grounded in open-access, evidence-based consumer health resources. The conversation was prompted to avoid speculative interpretation of mental states and instead focuses exclusively on observable routines.

Definition of all labelled routines (e.g., being still routine #2 is defined as more than 122.97 min being still between 12 am–9 am, no greater than 88.45 min being still between 9 am to 6 pm, no greater than 60.71 min after 6 pm) as well as the output of the GLM result are passed to the LLM.

Grouping Individuals based on GLM Results

Apart from the individuals with all anxiety or depression scores less than 3, we discovered from individual GLMs a list of significant routine types ($p < 0.05$) from any data categories. As there are more than 200 dominant routine types across all data categories and 215 participants, we chose to consider only the broad categories, removing routines from the same category and removing redundant categories that reflect similar aspects of life; for instance, the standard deviation and mean in amplitude audio are regarded as reflecting the same aspect of ambient noise. Concretely, mean coefficients were computed across routines from one sensing category, and the category with highest mean coefficients was retained among each of the category groups: ambient audio, containing (audio_amplitude_std, audio_amplitude_mean); conversation, containing (conversation_duration, conversation_num, audio_voice), phone use, containing (other_playing_duration, other_playing_num, unlock_duration, unlock_number), outgoing calls, containing (call_out_duration, call_out_num), incoming calls, containing (call_in_duration, call_in_num), light, containing (light_mean, light_std), distance travelled, containing (loc_dist, loc_max_dist_from_campus). Individuals were grouped if they shared the same set of categories. This resulted in 17 broader sensing categories, a reduction by 10 from the original 27 categories.

The previously retained list of individuals with no high (≥ 3) anxiety or depression scores was considered as one group. To assess whether there are significant between-group differences in anxiety, depression, and SSE scores, ANOVA was applied on GLMs built for anxiety and GLMs built for depression. Shapiro-Wilk and Levene tests were applied to assess normality and variance homogeneity to test the suitability of ANOVA. As the result of ANOVA is known to show the presence of between-group difference without the direction or magnitude of the difference, the non-anxiety and non-depression individuals would not be suitable to be included in ANOVA when assessing between-groups in anxiety and depression scores, but they were included to assess differences in SSE scores.

Data availability

All experiments were conducted on the open-access College Experience Study dataset. The dataset is planned to be released in stages. Up to the point this manuscript is submitted, only aggregated data has been released. The dataset is publicly available on Kaggle: <https://www.kaggle.com/datasets/subigyanepal/college-experience-dataset>.

Code availability

All scripts and pipelines used to process data and generate the results will be made publicly available upon publication at: <https://github.com/adrienschoi/npj-digmed-personalised-routine-variability>.

Received: 31 August 2024; Accepted: 26 August 2025;

Published online: 06 October 2025

References

- Price, G. D. et al. Using digital phenotyping to capture depression symptom variability: detecting naturalistic variability in depression

- symptoms across one year using passively collected wearable movement and sleep data. *Transl. Psychiatry* **13**, 381 (2023).
- Ratering, C., Van Der Heijden, R. & Martens, K. Moving around with an anxiety disorder. *Transp. Res. Part F. Traffic Psychol. Behav.* **100**, 493–506 (2024).
- Laiou, P. et al. The association between home stay and symptom severity in major depressive disorder: Preliminary findings from a multicenter observational study using geolocation data from smartphones. *JMIR Mhealth Uhealth* **10**, e28095 (2022).
- Bahri, A. A. et al. The relationship between walking and depression, anxiety, and stress among a sample from Jazan, Saudi Arabia: A cross-sectional investigation. *Medicine* **101**, e30718 (2022).
- Sabet, S. M., Dautovich, N. D. & Dzierzewski, J. M. The rhythm is gonna get you: social rhythms, sleep, depressive, and anxiety symptoms. *J. Affect. Disord.* **286**, 197–203 (2021).
- Althubaiti, A. Information bias in health research: definition, pitfalls, and adjustment methods. *J. Multidiscip. Healthc.* **9**, 211–217 (2016).
- Onnela, J.-P. & Rauch, S. L. Harnessing Smartphone-Based digital phenotyping to enhance behavioral and mental health. *Neuropsychopharmacology* **41**, 1691–1696 (2016).
- Huckvale, K., Venkatesh, S. & Christensen, H. Toward clinical digital phenotyping: a timely opportunity to consider purpose, quality, and safety. *NPJ Digit. Med.* **2**, 88 (2019).
- Eagle, N. & Pentland, A. S. Eigenbehaviors: identifying structure in routine. *Behav. Ecol. Sociobiol.* **63**, 1057–1066 (2009).
- Hou, W. K. et al. Measuring everyday processes and mechanisms of stress resilience: Development and initial validation of the Sustainability of Living Inventory (SOLI). *Psychol. Assess.* **31**, 715–729 (2019).
- Ehlers, C. L., Kupfer, D. J., Frank, E. & Monk, T. H. Biological rhythms and depression: The role of zeitgebers and zeitstörers. *Depress Anxiety* **1**, 285–293 (1993).
- Hou, W. K. et al. Regularizing daily routines for mental health during and after the COVID-19 pandemic. *J. Glob. Health* **10**, 020315 (2020).
- Liang, L. et al. Latent profiles and transitions of daily routine disruptions are associated with severity of symptoms of anxiety and depression. *Leis. Sci.* 1–16 <https://doi.org/10.1080/01490400.2023.2228311> (2023).
- Torous, J. et al. New Tools for New Research in Psychiatry: A scalable and customizable platform to empower data driven smartphone research. *JMIR Ment. Health* **3**, e16 (2016).
- Katsoulakis, E. et al. Digital twins for health: a scoping review. *NPJ Digit. Med.* **7**, 77 (2024).
- Loi, M. The digital phenotype: a philosophical and ethical exploration. *Philos. Technol.* **32**, 155–171 (2018).
- Mohr, D. C., Zhang, M. & Schueller, S. M. Personal sensing: Understanding mental health using ubiquitous sensors and machine learning. *Annu. Rev. Clin. Psychol.* **13**, 23–47 (2017).
- Bagot, K. S. et al. Current, future and potential use of mobile and wearable technologies and social media data in the ABCD study to increase understanding of contributors to child health. *Dev. Cogn. Neurosci.* **32**, 121–129 (2018).
- Jacobson, N. C., Summers, B. & Wilhelm, S. Digital Biomarkers of Social Anxiety Severity: Digital phenotyping using Passive smartphone sensors. *J. Med. Internet Res.* **22**, e16875 (2020).
- Boukhechba, M. et al. Monitoring social anxiety from mobility and communication patterns. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **1**, 1–24 (2017).
- Xu, X. et al. GLOBEM. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **6**, 1–34 (2022).
- DaSilva, A. W. et al. Correlates of stress in the college environment uncovered by the application of penalized generalized estimating equations to mobile sensing data. *JMIR Mhealth Uhealth* **7**, e12084 (2019).

23. Acikmese, Y. & Alptekin, S. E. Prediction of stress levels with LSTM and passive mobile sensors. *Procedia Comput. Sci.* **159**, 658–667 (2019).
24. Chikersal, P. et al. Detecting depression and predicting its onset using longitudinal symptoms captured by passive sensing: a machine learning approach with robust feature selection. *ACM Trans. Comput. Hum. Interact.* **28**, 1–41 (2021).
25. Rhim, S., Lee, U. & Han, K. Tracking and modeling subjective well-being using smartphone-based digital phenotype. *Proc. 28th ACM Conf. User Model. Adapt. Pers.* (2020).
26. Meyerhoff, J. et al. Evaluation of changes in depression, anxiety, and social anxiety using smartphone sensor features: Longitudinal cohort study. *J. Med. Internet Res.* **23**, e22844 (2021).
27. Xu, X. et al. Leveraging Routine Behavior and Contextually-Filtered Features for Depression Detection among College Students. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **3**, 1–33 (2019).
28. Zakaria, C., Balan, R. & Lee, Y. StressMon. *Proc. ACM Hum. -Comput. Interact.* **3**, 1–29 (2019).
29. Wang, R. et al. Tracking depression dynamics in college students using mobile phone and wearable sensing. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **2**, 1–26 (2018).
30. Liu, J. J. et al. Digital phenotyping from wearables using AI characterizes psychiatric disorders and identifies genetic associations. *Cell* **188**, 515–529 (2024).
31. Oudin, A. et al. Digital Phenotyping: Data-Driven Psychiatry to Redefine Mental Health. *J. Med. Internet Res.* **25**, e44502 (2023).
32. Nepal, S. et al. Capturing the college experience. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **8**, 1–37 (2024).
33. Kroenke, K., Spitzer, R. L., Williams, J. B. W. & Löwe, B. An Ultra-Brief screening scale for anxiety and depression: the PHQ-4. *Psychosomatics* **50**, 613–621 (2009).
34. Heatherton, T. F. & Polivy, J. Development and validation of a scale for measuring state self-esteem. *J. Pers. Soc. Psychol.* **60**, 895–910 (1991).
35. Wang, W. et al. First-Gen Lens: Assessing mental health of first-generation students across their first year at college using mobile sensing. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **6**, 95 (2022).
36. Obuchi, M. et al. Predicting brain functional connectivity using mobile sensing. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **4**, 1–22 (2020).
37. Huckins, J. et al. Mental Health and Behavior of College Students During the Early Phases of the COVID-19 Pandemic: Longitudinal Smartphone and Ecological Momentary Assessment Study. *J. Med. Internet Res.* **22**, e20185 (2020).
38. Rotter, J. B. Generalized expectancies for internal versus external control of reinforcement. *Psychol. Monogr.* **80**, 1–28 (1966).
39. Cho, Y.-J. et al. Integrative genomic analysis of medulloblastoma identifies a molecular subgroup that drives poor clinical outcome. *J. Clin. Oncol.* **29**, 1424–1430 (2011).
40. Barras, D. et al. BRAF V600E Mutant colorectal cancer subtypes based on gene expression. *Clin. Cancer Res.* **23**, 104–115 (2017).
41. Wang, P., Gao, L., Hu, Y. & Li, F. Feature related multi-view nonnegative matrix factorization for identifying conserved functional modules in multiple biological networks. *BMC Bioinformatics* **19**, 394 (2018).
42. Kleiman, E. M. et al. Digital phenotyping of suicidal thoughts. *Depress Anxiety* **35**, 601–608 (2018).
43. Goldstein, B. L. & Klein, D. N. A review of selected candidate endophenotypes for depression. *Clin. Psychol. Rev.* **34**, 417–427 (2014).
44. Gottesman, I. I. & Gould, T. D. The endophenotype Concept in Psychiatry: Etymology and Strategic Intentions. *Am. J. Psychiatry* **160**, 636–645 (2003).
45. Khasentino, J. et al. A personal health large language model for sleep and fitness coaching. *Nat. Med.* <https://doi.org/10.1038/s41591-025-03888-0> (2025).
46. Xu, X. et al. Mental-LLM: Leveraging Large Language Models for Mental Health Prediction via Online Text Data. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **8**, 31 (2024).
47. Kruzan, K. P. et al. The Perceived Utility of Smartphone and Wearable Sensor Data in Digital Self-tracking Technologies for Mental Health. *Proc. 2023 CHI Conf. Hum. Factors Comput. Syst.* **88** (2023).
48. Pataranutaporn, P. et al. Future You: A Conversation with an AI-Generated Future Self Reduces Anxiety, Negative Emotions, and Increases Future Self-Continuity. *arXiv* 2405.12514 (2024).
49. Fang, C. M. et al. PhysioLLM: Supporting Personalized Health Insights with Wearables and Large Language Models. IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI ...) 1–8 (2024) <https://doi.org/10.1109/bhi62660.2024.10913781>.
50. Ashman, S. B. et al. Relationship between social rhythms and mood in patients with rapid cycling bipolar disorder. *Psychiatry Res.* **86**, 1–8 (1999).
51. Boland, E. M. et al. Life events and social rhythms in bipolar spectrum disorders: An examination of social rhythm sensitivity. *J. Affect. Disord.* **139**, 264–272 (2012).
52. Grandin, L. D., Alloy, L. B. & Abramson, L. Y. The social zeitgeber theory, circadian rhythms, and mood disorders: Review and evaluation. *Clin. Psychol. Rev.* **26**, 679–694 (2006).
53. Mackett, R. L. Mental health and travel behaviour. *J. Transp. Health* **22**, 101143 (2021).
54. Choi, A., Ooi, A. & Lottridge, D. Digital Phenotyping for Stress, Anxiety, and Mild Depression: Systematic Literature review. *JMIR Mhealth Uhealth* **12**, e40689 (2024).
55. Kandola, A. et al. Physical activity and depression: Towards understanding the antidepressant mechanisms of physical activity. *Neurosci. Biobehav. Rev.* **107**, 525–539 (2019).
56. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders, 5th edn (American Psychiatric Publishing, 2013). <https://doi.org/10.1176/appi.books.9780890425596>.
57. Richardson, L. P. et al. Evaluation of the PHQ-2 as a Brief Screen for Detecting Major Depression Among Adolescents. *Pediatrics* **125**, e1097–e1103 (2010).
58. Generalized Anxiety Disorder 2-item (GAD-2). *National HIV Curriculum* <https://www.hiv.uw.edu/page/mental-health-screening/gad-2> (2025). University of Washington.

Acknowledgements

Not applicable.

Author contributions

A.C. conceived the study and conducted all experiments, the writing, supervised by J.W. (main supervisor for this study, who provided feedback on experimentation and the manuscript) and D.L. (provided feedback on the manuscript).

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Adrien Choi.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025