Proceedings of the 4th International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/2026.39483

Longitudinal Behavioral Sensing in Mental Health: Modelling Approaches, Methodological Challenges, and Foundation-Model Opportunities

Yi Zhou

Department of Mathematics, Imperial College London, London, UK
yz9423@ic.ac.uk

Abstract. Longitudinal behavioral sensing is increasingly used in mental-health monitoring and prediction.This review examines how longitudinal behavioral data are modelled in mental health, identifies methodological challenges and residual gaps after considering existing responses, evaluates the potential of foundation and pretrained models. Existing studies differ in their temporal representations, prediction horizons, personalization strategies, and evaluation settings. Common challenges include missingness in data, limited temporal coverage, sparse ground truth of mental health, individual heterogeneity, and generalization across sensing and study settings. Current approaches address parts of these challenges, while residual gaps remain. The unsolved challenges are particularly in preserving clinically relevant temporal information, adapting to individual behavioral patterns, also transferring models across different contexts. Foundation and pretrained models may offer potential solutions through learned temporal representations, learning from unlabeled sensing data, more efficient adaptation of data, and transferring across heterogeneous settings. Future research could adapt these capabilities to the longitudinal mental health context, particularly in modelling longer individual trajectories, enabling more effective personalization, and generalization across diverse populations and sensing contexts.

Keywords: longitudinal behavioral sensing, mental health, digital phenotyping, foundation models

1. Introduction

With the development of wearable devices and mobile sensing technology, it is now relatively easy to collect behaviour- and physiology-related signals continuously and passively in real-world scenarios. Repeated collection of these signals can generate longitudinal behavioural data. Such data collect many information about people's lives, such as physical activity, movement, sleep, social life, use of devices, etc., and have been linked to all kinds of health problems. Longitudinal behavioural data in this study refer to multiple or extended observations of a person's behaviour and physiological conditions taken at different times to investigate changes in behaviour and the reasons for such changes [1].

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.39483

Patterns of daily behaviour may be linked to mental health problems, and thus, research can be conducted on how long-term changes in behaviour relate to mental health. Continuous behavioural observation can be used to collect extended data on behaviour and changes in the long run. Supplement the clinical interview and self-report questionnaire data. With the accumulation of longitudinal behavioural data, many researchers now conduct studies to observe changes in anxiety and depression over time [1].

However, there are some problems with the behaviour-sensing data. Behavioural sensing data are often multi-modal and collected at different times [2]. Behavioral Patterns and Sensor-derived Features Also Differ Significantly [3]. In addition, the real-world sensing data are often incomplete due to both technical reasons and changes in behaviour [4, 5]. The time resolution of behavioural sensing data may also be different from that of mental health examination [6]. Therefore, it is not suitable for addressing the problems of time-series data, individual behaviour modelling and connecting sensor data to mental health.

Various Methods have been employed by researchers to model longitudinal behavioural data. The existing methods of temporal information processing are diverse, such as aggregated or manually engineered features in traditional machine learning and deep sequence models, and classical time-series models [4, 5, 7]. Collection of long-term data does not require the explicit modelling of temporal dependence. Observation summaries can be performed using the aforementioned temporal features. There are many types of methods, such as population-level models, localised and individualised ones, etc.; in addition, there are differences in input windows, prediction horizons and evaluation settings [2-8].

Foundation models are trained on vast amounts of data to learn general-purpose representations. Then, fine-tune or probe them for the downstream task using other transfer methods. This way has been extended to general time series modelling, wearable and behavioural sensing, and pretrained models have been developed for large-scale physiological, activity and behaviour data [9, 10].

Researchers have applied a range of approaches to model longitudinal behavioral data. Existing methods represent temporal information in different ways, including aggregated or manually engineered features used in conventional machine learning, deep sequence models, and classical time-series models [4, 5, 7]. Importantly, collecting longitudinal data does not necessarily mean that temporal dependencies are explicitly modelled. Since observations may be summarized using predefined temporal features instead. Approaches also vary from population-level models to locally personalized and individualized models, with further differences in input windows, prediction horizons, and evaluation settings [2, 3, 8].

Foundation models are typically pretrained on large scale datasets to learn generalizable representations. Then adapt them to downstream tasks through fine tuning, probing, or other transfer strategies. This approach has expanded from general time series modelling to wearable and behavioral sensing, where pretrained models have been developed using large scale physiological, activity, and behavioral data [9, 10].

Pre-trained and foundation-model approaches may help to address some of the methodological problems in longitudinal behaviour sensing. Self-supervised pre-training can learn representations from a large amount of unlabeled sensor data to reduce the need for task-specific labels [9, 11]. Recently, research has shown that the representations learned by wearable foundation models can be transferred to various sensing tasks and modalities [9, 12]. Evidence in mental health applications is emerging, but limited labelled data have been used [12]. However, research in this area is relatively scarce compared with other wearable and behaviour-sensing applications. However, pre-training cannot learn about truly unseen behaviour or reduce the demand for mental health ground truth.

DOI: 10.54254/2755-2721/2026.39483

Furthermore, the effect of individual-level personalisation and generalisation across populations, devices, and datasets has not been well demonstrated in mental health sensing.

Previous studies have explored the use and methods of digital phenotyping and passive sensing in mental health. Longitudinal mental health monitoring has also been studied [1]. However, several questions remain unaddressed. First, the existing research has not sufficiently reported on how longitudinal behavioural data are presented and modelled in the studies, and whether temporal information is explicitly preserved and modelled or summarised through aggregated features. Second, in the study of methodological deficiencies in existing solutions, some solutions are constrained. Comparison of the modelling and evaluation of multiple studies may also reveal deficiencies that have gone unnoticed in individual research. Third, it is unknown whether and how the capabilities of foundation models in other sensing fields can address the deficiencies in longitudinal mental health monitoring. Therefore, this paper will explore how longitudinal behavioural data are modelled in the field of mental health, identify the methodological deficiencies of current studies, and examine the prospects of pretrained and foundation models in addressing these deficiencies. Distinguish between the established evidence and the new opportunities or problems in this study.

2. Related work

This section reviews how longitudinal behavioral information has been represented and modelled in mental-health sensing studies. The discussion covers existing modelling and representation-learning approaches, while foundation models and their potential implications are examined separately in Section 4. The literature is organized according to how longitudinal observations are transformed into model inputs, how temporal information is represented, the temporal relationship between inputs and outcomes, approaches to personalization, and evaluation settings.

2.1. From longitudinal observation to model inputs

Observation periods vary across longitudinal behavioral studies, from days to months or longer. Studies also differ in aggregation window, which is how observations are aggregated over time, and the training history. It is important to distinguish the overall observation span from the model input window, as they may differ substantially.

These temporal scales can differ within the same study. For example, Aalbers et al. collected repeated observations over a period of up to 60 days, but each stress prediction mainly used behavior from the preceding 60 minutes [13]. Similarly, Kathan et al. followed participants for approximately eight weeks, while each model input contained up to seven days of sequential data [14]. In terms of data aggregation, Langener et al. used aggregation windows of 3, 6, and 24 hours within a 28 day study [2].

2.2. Representation of temporal information

Studies have taken different approaches to modelling temporal information, including feature based methods, sequence models, and classical time-series models. Feature based approaches summarize repeated observations into predefined temporal features before modelling. Chikersal et al. constructed weekly, accumulated, and behavioral change features from passive sensing data [4]. Jang et al. used 24 hour digital phenotype features [15], while Lipschitz et al. summarized Fitbit

DOI: 10.54254/2755-2721/2026.39483

measures over two week windows [16]. In these studies, temporal information is mainly represented through aggregation and engineered features

Sequence models preserve the order of repeated observations and learn temporal dependencies during training. Kathan et al. used a Gated Recurrent Unit (GRU) with sequences of up to seven days for depressive-symptom prediction [14]. Zou et al. used daily sequences with Long Short Term Memory (LSTM), GRU, and Gated Recurrent Unit with Decay (GRU-D). GRU-D also included missingness masks and the time since the previous observation [5].

Classical time-series models represent temporal structure explicitly. Yokoyama et al. converted accelerometer data into hourly physical-activity time series and fitted an individual Seasonal Autoregressive Integrated Moving Average (SARIMA) model with 24-hour seasonality for each participant [7].

2.3. Temporal relationship between input and outcome

Studies differ in the temporal relationship between behavioral inputs and mental-health outcomes. These tasks can be grouped into concurrent estimation, short term forecasting, and longer-term prediction. Concurrent estimation links behavioral data to outcomes from the same or corresponding period. Aalbers et al. used behavior preceding each Ecological Momentary Assessment (EMA) to estimate momentary stress [13]. Lipschitz et al. summarized Fitbit measures over the same two-week periods as depressive and hypomanic/manic symptoms [16].

Short-term forecasting uses recent behavioral data to predict a subsequent mental-health state. Jang et al. used 24-hour behavioral features to identify impending panic symptoms on the following day [15]. Longer-term prediction uses earlier behavioral data to predict outcomes several weeks later. Zou et al. used early-treatment sequences to predict treatment response approximately 8–10 weeks before treatment completion [5]. Chikersal et al. predicted later depression and symptom worsening at successive semester cutoffs, in some cases around 11–15 weeks before semester end [4]. The prediction horizon is distinct from the input window in Section 2.1, which describes how much previous behavioral data are used for each prediction.

2.4. Levels of personalization

Studies incorporate individual differences at different levels. Population-level models use data from multiple participants to learn a common relationship between behavioral patterns and mental-health outcomes. Person-aware or locally personalized approaches retain a shared model but incorporate individual information or select more relevant training data for each participant. Fully individualized models instead learn behavioral relationships separately for each person. Another approach combines shared learning across participants with participant-specific adaptation. Phanord et al. compared group-general, group-personalized, and within-person models [8], while Adler et al. used behaviorally similar participants for local personalization [3]. Shah et al. trained separate models for each participant [17], whereas Kathan et al. combined shared and participant-specific model components [14].

Within-person analysis does not necessarily mean that a model is personalized. Lakhtakia et al., for example, separated within- and between-person associations without training participant-specific predictive models [6].

DOI: 10.54254/2755-2721/2026.39483

2.5. Evaluation and generalization

Studies differ in both the metrics used to report performance and the temporal tasks being evaluated. Classification studies use measures such as accuracy and Area Under the Receiver Operating Characteristic Curve (ROC-AUC) [4, 15], while continuous-outcome studies may instead use error-based measures such as Mean Absolute Error (MAE) [14]. Temporal tasks also range from concurrent estimation to next-day forecasting and prediction several weeks in advance [4, 13, 14].

Studies use different evaluation settings to test generalization. Within-person or temporal evaluation uses new observations from previously observed participants, while participant-level splits evaluate unseen participants. Cross-dataset evaluation tests generalization across different datasets. For example, Adler et al. used leave-one-subject-out validation and cross-dataset evaluation [3]. In contrast, Kathan et al. included the same participants during model development and testing [14].

3. Limitation, response and residual

3.1. Data acquisition and availability

Real world behavioral sensing data are often affected by missingness and measurement noise. Missing data can arise from both technical problems and user-related factors, which may also be difficult to distinguish. For example, Zou et al. reported missingness associated with technical and user-related causes [5], while Kathan et al. identified disruptions from participants' own phones and applications [14]. Similarly, Chikersal et al. identified missing and noisy data as challenges in passive-sensing studies [4]. Noisy or inconsistent measurements can also make it difficult to distinguish real behavioral changes from variation introduced during data collection.

Missingness and low quality data can substantially reduce the amount of usable data. In Kathan et al., only 16 of 65 recruited participants were retained for the main modelling analyses after quality control [14]. Zou et al. also found that filtering out participants with insufficient data reduced the sample size for each prediction task [5].

Such data loss further reduces the sample sizes available for modelling. Although repeated sensing produces many observations, the number of independent participants can remain limited. This could be reflected in Langener et al. [2], Shah et al. [17], and Phanord et al. [8]. Consequently, more repeated measurements do not necessarily provide greater population coverage.

Existing approaches can accommodate missingness through filtering, imputation, or missingness-aware modelling, but cannot recover behavior that was never observed. Moreover, stricter data-quality requirements may further reduce the effective sample size. Data availability therefore remains closely coupled with the amount and representativeness of information available for modelling.

3.2. Temporal representation and coverage

Aggregating longitudinal behavioral data can result in the loss of temporal information. Zou et al. noted that simple statistical operations can lose temporal tendency information that may be important for treatment-response prediction [5]. Yokoyama et al. similarly identified temporal aggregation as a limitation after converting 20-Hz accelerometer data into minute-level activity intensity and then hourly totals [7]. This type of aggregation may obscure the temporal order of observations and changes in behavior over time.

DOI: 10.54254/2755-2721/2026.39483

Long observation periods also do not necessarily correspond to long model input windows. As a result, behavioral patterns outside the model input window may not be directly represented in each prediction. Models may therefore capture short-term behavioral patterns while missing changes which develop over longer periods.

Although temporally informed features and sequence models can preserve more longitudinal structure, they capture behavioral patterns at different temporal resolutions and time scales. It remains unclear which clinically relevant longitudinal patterns are preserved or overlooked by these representations.

3.3. Mental-health ground truth

Behavioral sensing data and mental-health assessments are often collected at different resolutions. Behavioral signals can be collected continuously, whereas mental-health outcomes are only available at particular assessment points. For example, Lipschitz et al. collected Fitbit data over nine months, but mood symptoms were assessed every two weeks [16]. Lakhtakia et al. collected passive smartphone data throughout a 16-week study, while symptom assessments using EMA were conducted only during selected periods [6]. Chikersal et al. reported a similar mismatch between longitudinal sensing and less frequent depression assessments [4]. As a result, behavioral changes may be recorded much more frequently than changes in mental-health symptoms, making it difficult to determine how the two are related over time.

The measured outcome may also capture only part of the mental health condition. Depression includes different symptoms, which may be associated with different behavioral patterns.. For example, Lakhtakia et al. reported different associations for affect, stress, fatigue, and distractibility, and noted that the selected EMA measures did not cover the full heterogeneity of depression [6]. Using a single overall severity measure may therefore hide some of these symptom specific relationships.

Repeated assessments and symptom specific outcomes have partly addressed these limitations by providing denser and more granular ground truth. However, assessments typically remain much sparser than continuous sensing, while selected symptom measures may still capture only part of the clinical construct. Ground truth therefore remains a constraint on longitudinal modelling that improvements in behavioral representation alone cannot resolve.

3.4. Individual heterogeneity and personalization

Between-person heterogeneity is common in behavioral–mental-health relationships. The behavioral features associated with mental-health outcomes, and the temporal scales at which these relationships appear, can differ across individuals. Langener et al. found that both the useful predictors and the relevant temporal scales differed between participants [2]. This heterogeneity creates different limitations depending on how individual differences are represented in the model.

Personalization has been used to account for such heterogeneity, but it introduces a trade-off between shared learning and individual specificity. Population-level models may fail to capture person-specific relationships, as suggested by differences in predictive performance between group-general and more personalized models in [8]. Greater personalization can reduce assumptions of shared relationships but also reduces the amount of data available for learning; Aalbers et al. identified limited observations per participant as a constraint on idiographic modelling [13]. Moreover, person-aware models may capture stable participant identity rather than dynamic

DOI: 10.54254/2755-2721/2026.39483

behavioral–symptom relationships. The remaining challenge is therefore to model individual heterogeneity without sacrificing the data sharing needed for robust learning and generalization.

3.5. Generalization and evaluation

Differences across studies can limit the transfer of behavioral sensing models. Adler et al. compared two datasets that differed in their populations, study periods, sensing systems, and available variables [3]. Only features available in both datasets could be aligned for cross-study modelling, and combining the datasets did not consistently improve generalization. Zou et al. also identified device, application, and study-specific sensing settings as limitations to generalizability [5]. Differences in sensing systems can therefore affect which behavioral variables are available across datasets and what information can be transferred between them. Lakhtakia et al. also noted that their study was limited to Android devices and included relatively few participants from racial and ethnic minority groups [6]. Differences in population and study context may also make one group or setting transferring to another group difficult.

Different evaluation settings also answer different generalization questions. Performance on future observations from known participants does not show whether the same model will work for a new participant. Additionally, performance on unseen participants within one dataset does not establish generalisation to a different dataset or sensing setting. In practice, however, evaluation settings often vary together with dataset boundaries. This makes it difficult to determine which differences contribute to changes in model performance.

Differences in prediction tasks and evaluation metrics also make it difficult to compare results across studies. Studies do various tasks including concurrent estimation, next day forecasting, or prediction several weeks in advance, while studies also use different task-specific performance measures. As a result, higher reported performance does not necessarily indicate a better model, since prediction difficulty and evaluation methods may differ.

4. Foundation-model opportunities for mental-health behavioral sensing

The following discussion synthesizes capabilities demonstrated across different studies rather than directly comparing their performance, as these studies differ in their tasks, temporal settings, evaluation metrics, and evaluation designs.

4.1. Data acquisition and availability

Foundation models partially address missing-data limitations by making useful representation learning less dependent on complete observations. Their main contribution is therefore not to eliminate missingness, but to reduce the extent to which incomplete observations prevent behavioral representations from being learned. This is already demonstrated in mental-health sensing through masked pretraining on partially observed actigraphy in Pretrained Actigraphy Transformer (PAT), while adjacent foundation models extend this capability to noisy and irregular behavioral time series and sensor imputation [9, 11, 12]. These findings suggest that representation learning can remain useful under forms of incompleteness common in real-world behavioral sensing. The remaining limitation is narrower: robustness to incomplete inputs does not recover genuinely unobserved behavior or resolve informative real-world missingness.

Foundation models provide a clearer response to label scarcity than to participant scarcity. Pretraining allows behavioral representations to be learned from large quantities of unlabeled data

DOI: 10.54254/2755-2721/2026.39483

and subsequently reused with fewer task-specific labels, as demonstrated in mental-health-relevant tasks by PAT and in adjacent behavioral sensing through self-supervised transfer learning. This could reduce the dependence of mental-health models on densely labelled longitudinal datasets. It does not, however, increase the number or diversity of individuals represented. Demographic underrepresentation remains even in Wearable Behavior Model (WBM) 's large-scale wearable pretraining cohort [10].

4.2. Temporal representation and coverage

Foundation models partly address limitations in temporal representation by learning representations directly from longitudinal sequences rather than first reducing them to a small set of handcrafted statistics. PAT demonstrates this directly in mental-health sensing. Week-long minute-level actigraphy is divided into temporal patches and partially masked, allowing a Transformer encoder to learn representations from the remaining sequence through masked-autoencoder pretraining before transfer to mental-health-related tasks [12]. Adjacent behavioral foundation models demonstrate that learned temporal representations can also be developed at larger scale. Joint Embedding for Time Series (JETS) learns latent representations from the temporal context of irregular behavioral observations [11], while WBM learns weekly behavioral embeddings from large-scale wearable data [10]. Rather than relying entirely on manually defined temporal features, these models learn representations from behavioral data. Nevertheless, the resulting representations remain forms of compression, and their use does not guarantee that all clinically relevant temporal structure is preserved.

Temporal coverage remains a challenge. Training on larger datasets does not necessarily mean that a model can use longer individual histories when making predictions. WBM uses one-week modelling windows despite being pretrained on approximately 2.5 billion hours of data [10], while Large Sensor Model (LSM) uses 300-minute samples despite being pretrained on a corpus of approximately 40 million hours [9]. The challenge is therefore to extend representation learning across longer individual histories rather than simply across more short observations. Adjacent wearable foundation model research has proposed hierarchical temporal representations and retrieval based personal histories for this purpose. Applying these approaches to mental health sensing could allow information across days, weeks, or longer periods to be integrated without requiring the entire high resolution history to be processed as a single sequence.

Extending the prediction horizon is another potential direction for mental health foundation models. Longer historical context does not itself enable prediction further into the future, and current mental health foundation models provide limited evidence for long-horizon forecasting. Adjacent foundation models, however, demonstrate that pretrained temporal models can extrapolate beyond observed windows. Large Sensor Model (LSM) forecasts future wearable signals [9], while Medical Foundation Model for Irregularity Adaptation (MIRA) forecasts irregular medical time series at arbitrary target timestamps through continuous latent-trajectory modelling [18]. These capabilities have not yet been established for clinically meaningful mental-health forecasting but suggest a direction beyond concurrent symptom estimation. Future mental-health foundation models could combine long-term personal representations with explicit forecasting objectives to predict changes in symptoms or episode risk over clinically relevant horizons.

DOI: 10.54254/2755-2721/2026.39483

4.3. Mental-health ground truth

Foundation models change the dependence of mental-health sensing on sparse ground truth. Self-supervised pretraining allows behavioral representations to be learned from largely unlabeled sensing data, leaving sparse mental-health labels mainly to anchor them to downstream outcomes. This shift is already evident in PAT and is supported by transfer learning under limited downstream labels [12]. However, reducing the need for labelled training data does not resolve the limited temporal coverage of mental health assessments. For instance, NarrativeSense predicts weekly affect scores derived from self reported assessments, which remain necessary to define and evaluate the prediction target [19]. Foundation models could also be used to identify informative periods for additional ecological momentary assessment or clinical assessment, moving from tolerating sparse supervision toward adaptive ground-truth collection.

A second opportunity is to broaden the evidence used to represent mental-health states. The value of multimodal foundation models lies not simply in processing more signals but in integrating complementary information. Adjacent wearable foundation models already demonstrate this capability. Heterogeneous physiological signals can be jointly represented, while combining behavioral and photoplethysmography representations improves health prediction [9, 10]. For mental health, complementary measurements may therefore be suggested, integrating behavioral trajectories with physiological signals, repeated symptom reports, and relevant contextual or clinical information according to what each contributes. Foundation models could provide a shared representation for these heterogeneous observations, although they cannot recover information absent from all included measurements.

Foundation models may also change how behavioral information is interpreted. NarrativeSense demonstrates that smartphone derived behavioral patterns can be contextualized as everyday experiences, while adjacent sensor language models show that sensor representations can be aligned with semantic information [19]. The opportunity is not narrative generation itself but making behavioral meaning conditional on information beyond the current sensing window. This suggests an opportunity for context conditioned longitudinal interpretation, combining current behavioral changes with personal baselines, previous symptom trajectories, and current context. The same behavioral change could then carry different meaning across individuals or circumstances, while contextual interpretation remains distinct from causal explanation.

4.4. Individual heterogeneity and personalization

Foundation models may shift the trade-off between population-level learning and individual personalization by separating the learning of shared behavioral structure from adaptation to a specific person. Population models benefit from larger pooled datasets, whereas person-specific models capture individual variation but rely on much less data. Pretraining offers a possible middle ground. Shared behavioral representations can first be learned from large populations and then adapted using less downstream data, allowing individual modelling to start from a population-informed representation rather than limited personal observations alone. This is partly supported in mental health by PAT and in adjacent wearable sensing by NormWear, which demonstrates adaptation from pretrained representations with limited downstream data [12, 20]. Foundation-model pretraining therefore provides a potential population prior for more data-efficient personalization. Current evidence, however, mainly demonstrates adaptation across tasks or datasets rather than adaptation to an individual's behavioral baseline.

DOI: 10.54254/2755-2721/2026.39483

Foundation models could move beyond data-efficient adaptation by conditioning pretrained representations on individual behavioral history. Population-level knowledge could be retained while current representations are adjusted according to the individual's previous patterns. The next step is to extend this adaptation from selected historical observations to evolving longitudinal trajectories, as also proposed in recent wearable foundation-model research. Therefore, this suggests population-pretrained, person-conditioned modelling that integrates shared representations with individual baselines, accumulated history, and recent trajectories for continual personalization over time.

4.5. Generalization and evaluation

Foundation models may improve cross-study transfer by reducing dependence on manually aligned feature spaces. Conventional transfer can require datasets to be restricted to overlapping engineered features, as in [3], potentially discarding study-specific information before modelling. In contrast, adjacent wearable foundation models demonstrate that reusable representations can be learned across heterogeneous sensor configurations and transferred across downstream settings. NormWear, for example, supports varying multivariate inputs and zero-, few-, and full-shot transfer across multiple health tasks [20]. This suggests that foundation models may reduce representation incompatibility across sensing settings. Evidence that they resolve broader population- or context-dependent shifts in behavioral–mental-health relationships remain limited.

Foundation models do not directly resolve differences in evaluation settings. Zero, few, and full-shot adaptation instead introduce additional dimensions along which generalization claims can differ. Adjacent foundation-model research similarly emphasizes that robustness should be evaluated under the specific dataset, device, or user shifts being claimed. Future mental-health foundation-model studies should therefore report temporal task definition, prediction horizon, participant split, dataset or device shift, and adaptation regime together so that reported generalization performance can be interpreted and compared consistently.

5. Conclusion

This review examined how longitudinal behavioral data are modelled in mental health sensing, identified methodological challenges and residual gaps after considering existing responses. It evaluated the opportunities offered by foundation and pretrained models. Current studies varies in how temporal information is represented, how individual differences are incorporated, and how predictive performance is evaluated. Although existing approaches have addressed parts of these challenges, limitations remain in temporal coverage, sparse mental health ground truth, individual heterogeneity, and generalization across sensing settings. These challenges are also interconnected. Data availability and temporal representation influence which behavioral patterns a model can capture, while evaluation design affects how the performance is assessed.

Pretrained and foundation models may help address these challenges by learning temporal representations, reducing reliance on task specific labels, and supporting adaptation across different tasks and sensing settings. However, pretraining cannot recover unobserved behavior or missing mental health assessments. Moreover, improvements in representation learning do not necessarily result in better long term forecasting, individual personalization, or cross study generalization. Therefore, their potential depends on how pretrained representations are adapted to the temporal and individual characteristics of the specific data.

DOI: 10.54254/2755-2721/2026.39483

Future research could investigate incorporating longer individual histories into modelling, clinically meaningful forecasting, personal adaptation, and transfer across populations, devices, and study settings. In particular, opportunities lie both in transferring existing pretrained models to mental health applications and in developing foundation models that better accommodate longitudinal and behavioral patterns. Further evaluation across different temporal tasks, populations, and sensing settings could help determine which methodological gaps these approaches can address and which limitations remain unresolved.

References

[1] Amin, R., Schreynemackers, S., Oppenheimer, H., Petrovic, M., Hegerl, U., & Reich, H. (2025). Use of mobile sensing data for longitudinal monitoring and prediction of depression severity: Systematic review. *Journal of Medical Internet Research, 27*, e57418. https://doi.org/10.2196/57418

[2] Langener, A. M., Bringmann, L. F., Kas, M. J., & Stulp, G. (2024). Predicting mood based on the social context measured through the experience sampling method, digital phenotyping, and social networks. *Administration and Policy in Mental Health and Mental Health Services Research, 51*, 455–475. https://doi.org/10.1007/s10488-023-01328-0

[3] Adler, D. A., Wang, F., Mohr, D. C., & Choudhury, T. (2022). Machine learning for passive mental health symptom prediction: Generalization across different longitudinal mobile sensing studies. *PLOS ONE, 17*(4), e0266516. https://doi.org/10.1371/journal.pone.0266516

[4] Chikersal, P., Doryab, A., Tumminia, M., Villalba, D. K., Dutcher, J. M., Liu, X., Cohen, S., Creswell, K. G., Mankoff, J., Creswell, J. D., Goel, M., & Dey, A. K. (2021). Detecting depression and predicting its onset using longitudinal symptoms captured by passive sensing: A machine learning approach with robust feature selection. *ACM Transactions on Computer-Human Interaction, 28*(1), Article 3. https://doi.org/10.1145/3422821

[5] Zou, B., Zhang, X., Xiao, L., Bai, R., Li, X., Liang, H., Ma, H., & Wang, G. (2023). Sequence modeling of passive sensing data for treatment response prediction in major depressive disorder. *IEEE Transactions on Neural Systems and Rehabilitation Engineering, 31*, 1786–1795. https://doi.org/10.1109/TNSRE.2023.3260301

[6] Lakhtakia, T., Smith, S. R., Meyerhoff, J., Meng, Y., Ungar, L. H., Mohr, D. C., & Stamatis, C. A. (2025). Differential associations of passively sensed behaviors with in-vivo depression symptoms. *Journal of Psychopathology and Clinical Science*. https://doi.org/10.1037/abn0001059

[7] Yokoyama, S., Kagawa, F., Takamura, M., Takagaki, K., Kambara, K., Mitsuyama, Y., Shimizu, A., Okada, G., & Okamoto, Y. (2023). Day-to-day regularity and diurnal switching of physical activity reduce depression-related behaviors: A time-series analysis of wearable device data. *BMC Public Health, 23*, Article 34. https://doi.org/10.1186/s12889-023-14984-6

[8] Phanord, C. S., Ruzic, L. L., Kalyanasundaram, S., Barnes-Horowitz, S., Friedman, N. P., Chaspari, T., & Kaiser, R. H. (2026). Digital phenotyping of affect and stress in emerging adults. *Frontiers in Digital Health, 8*, Article 1799541. https://doi.org/10.3389/fdgth.2026.1799541

[9] Narayanswamy, G., Liu, X., Ayush, K., Yang, Y., Xu, X., Liao, S., Garrison, J., Tailor, S., Sunshine, J., Liu, Y., Althoff, T., Narayanan, S., Kohli, P., Zhan, J., Malhotra, M., Patel, S., Abdel-Ghaffar, S., & McDuff, D. (2025). Scaling wearable foundation models. *Proceedings of the 13th International Conference on Learning Representations (ICLR 2025)*.

[10] Erturk, E., Kamran, F., Abbaspourazad, S., Jewell, S., Sharma, H., Li, Y., Williamson, S., Foti, N. J., & Futoma, J. (2025). Beyond sensor data: Foundation models of behavioral data from wearables improve health predictions. *Proceedings of the 42nd International Conference on Machine Learning, 267*. https://proceedings.mlr.press/v267/erturk25a.html

[11] Xie, E., Rodriguez Martinez, R., Chang, W., & Ballinger, B. (2025). JETS: A self-supervised joint embedding time series foundation model for behavioral data in healthcare. [Preprint].

[12] Ruan, F. Y., Zhang, A., Oh, J. Y., Jin, S., & Jacobson, N. C. (2026). A foundation model for wearable movement data in mental health research. *IEEE Journal of Biomedical and Health Informatics*. https://doi.org/10.1109/JBHI.2026.3694809

[13] Aalbers, G., Hendrickson, A. T., Vanden Abeele, M. M. P., & Keijsers, L. (2023). Smartphone-tracked digital markers of momentary subjective stress in college students: Idiographic machine learning analysis. *JMIR mHealth and uHealth, 11*, e37469. https://doi.org/10.2196/37469

DOI: 10.54254/2755-2721/2026.39483

[14] Kathan, A., Harrer, M., Küster, L., Triantafyllopoulos, A., He, X., Milling, M., Gerczuk, M., Yan, T., Rajamani, S. T., Heber, E., Grossmann, I., Ebert, D. D., & Schuller, B. W. (2022). Personalised depression forecasting using mobile sensor data and ecological momentary assessment. *Frontiers in Digital Health, 4*, Article 964582. https://doi.org/10.3389/fdgth.2022.964582

[15] Jang, S., Sun, T. H., Shin, S., Lee, H.-J., Shin, Y.-B., Yeom, J. W., Park, Y. R., & Cho, C.-H. (2024). A digital phenotyping dataset for impending panic symptoms: A prospective longitudinal study. *Scientific Data, 11*, Article 1264. https://doi.org/10.1038/s41597-024-04147-6

[16] Lipschitz, J. M., Lin, S., Saghafian, S., Pike, C. K., & Burdick, K. E. (2024). Digital phenotyping in bipolar disorder: Using longitudinal Fitbit data and personalized machine learning to predict mood symptomatology. *Acta Psychiatrica Scandinavica*. https://doi.org/10.1111/acps.13765

[17] Shah, R. V., Grennan, G., Zafar-Khan, M., Alim, F., Dey, S., Ramanathan, D., & Mishra, J. (2021). Personalized machine learning of depressed mood using wearables. *Translational Psychiatry, 11*, Article 338. https://doi.org/10.1038/s41398-021-01445-0

[18] Li, H., Deng, B., Xu, C., Feng, Z., Schlegel, V., Huang, Y.-H., Sun, Y., Sun, J., Yang, K., Yu, Y., & Bian, J. (2025). MIRA: Medical time series foundation model for real-world health data. *Advances in Neural Information Processing Systems, 38*.

[19] Zhang, T., Li, Y., Ding, Y., Jia, H., Kostakos, V., & D'Alfonso, S. (2026). NarrativeSense: Predicting affective states in university students through smartphone sensing and contextual narratives. *ACM Transactions on Computing for Healthcare, 7*(2), Article 20. https://doi.org/10.1145/3788688

[20] Luo, Y., Chen, Y., Salekin, A., & Rahman, T. (2026). Toward foundation model for multivariate wearable sensing of physiological' signals. *ACM Transactions on Computing for Healthcare, 7*(3), Article 43. https://doi.org/10.1145/3803808