Beyond the Binary: Operationalizing Receptivity to Digital Health Interventions as a Time-to-Event Spectrum

SAMARTH NEGI*, Singapore-ETH Centre, Singapore
VARUN MISHRA, Northeastern University, United States
CHAI YIN KUM, National University of Singapore, Singapore
OSCAR CASTRO, Singapore-ETH Centre, Singapore
TOBIAS KOWATSCH, University of Zurich, Switzerland
JACQUELINE L. MAIR†, Singapore-ETH Centre, Singapore
FLORIAN VON WANGENHEIM†, Singapore-ETH Centre, Singapore

Just-In-Time Adaptive Interventions (JITAIs) aim to support health behavior by providing the right support at the right time. A critical determinant of JITAI efficacy is timing delivery such that the user is *receptive*, defined as the cognitive and behavioral capacity to receive, process, and use support. While prior work has explored context sensing to predict receptivity, standard approaches typically operationalize this construct as a binary outcome within a fixed window, despite theoretical definitions characterizing availability as a continuous, time-varying state. Modeling receptivity at this granularity increases learning complexity, and deep sequence models are further constrained by the scarcity of labeled interaction data in mHealth settings. To address these challenges, we propose `PRISM`, a deep learning framework for modeling receptivity as a probabilistic time-to-event distribution from longitudinal mobile sensing data. PRISM combines a Channel-Independent Transformer (PatchTST) encoder with a discrete-time survival objective and employs self-supervised pre-training on unlabeled sensor traces via Masked Patch Reconstruction to mitigate label scarcity.

We evaluate PRISM on the LvL UP intervention dataset, leveraging data from preliminary studies for pre-training and a large-scale efficacy trial for evaluation. Our results demonstrate competitive performance with established receptivity benchmarks, with self-supervised initialization yielding up to 15.5% improvement in median AUC across heterogeneous user groups. Beyond single-window evaluation, PRISM remains stable across multiple decision horizons from a single trained model, in contrast to binary baselines that degrade or require retraining at each cutoff. A follow-up evaluation on an independent dataset with variable prompt timing provides preliminary evidence that the learned representations transfer across cohorts and schedules. These findings suggest that PRISM provides a data-efficient pathway for deploying resilient, time-aware receptivity models in real-world mHealth systems.

†Co-author equal contribution.
*Corresponding author.

Authors' Contact Information: Samarth Negi (corresponding author), Singapore-ETH Centre, Future Health Technologies Programme, Singapore, Singapore, sanegi@student.ethz.ch; Varun Mishra, Northeastern University, Boston, Massachusetts, United States, v.mishra@northeastern.edu; Chai Yin Kum, National University of Singapore, School of Computing, Singapore, Singapore, kumchaiyin@u.nus.edu; Oscar Castro, Singapore-ETH Centre, Future Health Technologies Programme, Singapore, Singapore, oscar.castro@sec.ethz.ch; Tobias Kowatsch, University of Zurich, Institute for Implementation Science in Health Care, Zurich, Switzerland and University of St. Gallen, School of Medicine, St. Gallen, Switzerland and ETH Zurich, Department of Management, Technology, and Economics, Zurich, Switzerland, tobias.kowatsch@uzh.ch; Jacqueline L. Mair, Singapore-ETH Centre, Future Health Technologies Programme, Singapore, Singapore and National University of Singapore, Centre for Behavioural and Implementation Science Interventions, Yong Loo Lin School of Medicine, Singapore, Singapore, jacqueline.mair@sec.ethz.ch; Florian von Wangenheim, Singapore-ETH Centre, Future Health Technologies Programme, Singapore, Singapore and ETH Zurich, Department of Management, Technology, and Economics, Zurich, Switzerland, fwangenheim@ethz.ch.



This work is licensed under a Creative Commons Attribution 4.0 International License.
© 2026 Copyright held by the owner/author(s).
ACM 2474-9567/2026/9-ART141
https://doi.org/10.1145/3832017

CCS Concepts: • **Human-centered computing** → **Ubiquitous computing**; **Mobile computing**; Empirical studies in ubiquitous and mobile computing; • **Mathematics of computing** → **Survival analysis**; • **Computing methodologies** → *Neural networks*; *Unsupervised learning*; • **Applied computing** → *Health informatics*.

Additional Key Words and Phrases: just-in-time adaptive interventions, receptivity, mobile sensing, mHealth, time-to-event modeling, survival analysis, self-supervised learning, transformers, label efficiency

ACM Reference Format:
Samarth Negi, Varun Mishra, Chai Yin Kum, Oscar Castro, Tobias Kowatsch, Jacqueline L. Mair, and Florian von Wangenheim. 2026. Beyond the Binary: Operationalizing Receptivity to Digital Health Interventions as a Time-to-Event Spectrum . *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 10, 3, Article 141 (September 2026), 27 pages. https://doi.org/10.1145/3832017

1 INTRODUCTION

Just-In-Time Adaptive Interventions (JITAIs) aim to deliver support that is temporally precise and contextually relevant [12, 48]. JITAI frameworks are grounded in the idea that timing is identified through 'states of vulnerability and opportunity'; they aim to provide support at moments of need (e.g., low self-regulation) or opportunity (e.g., a moment to act positively in line with goals) [12, 39]. However, even when a state of need is correctly identified, the intervention may fail if delivered when the user is unable to attend to it, potentially causing frustration or burden[15]. To ensure support is delivered when it can be used, researchers are actively testing the dimension of receptivity, defined as the transient state of a user's ability and willingness to receive, process, and use support [32]. Determining exactly how to measure this capacity remains a complex challenge[17, 42]. Receptivity is currently treated as a binary outcome (receptive/not receptive), despite being theoretically established as a continuous construct governed by context and intervention burden.

Advances in mobile sensing and wearable technology have opened up the possibility of adapting intervention delivery to a user's changing internal state and external context[46, 49]. By continuously tracking physiological and behavioral markers, these technologies enable JITAIs to potentially detect the precise moments when a user is most able to engage with support[29]. Yet, a critical disconnect remains between the richness of this sensing data and the coarseness of current modeling. In practice, receptivity is operationalized through a fixed time-window threshold: a user is labeled 'receptive' (y=1) only if they interact with a prompt within a specific duration (e.g., 10 minutes) and 'non-receptive' (y=0) otherwise[17, 29]. While this abstraction was instrumental in establishing that mobile sensing could predict response probabilities better than chance, it inherently flattens the rich temporal dynamics of the user's context. By enforcing a rigid cutoff, binary models artificially conflate slightly delayed responses with total non-engagement, creating a 'noisy negative' target that obscures the nuance between immediate interruptibility and deferred availability. To better align measurement with the theoretical definition of receptivity as a continuous spectrum, we propose a time-to-event framework. Unlike binary classifiers that force a decision based on an arbitrary timeout, a time-to-event model predicts the distribution of likely response times, allowing JITAIs to optimize for the user's transient ability to receive, process, and use support.

Deep sequence models have revolutionized temporal modeling in various mobile sensing domains, ranging from Human Activity Recognition[8, 9] to disease progression monitoring[50], their application to behavioral outcomes, including detecting receptivity, remains largely underexplored. Current JITAI systems predominantly rely on static decision rules or basic classifiers[47], failing to capture the long-term fine-grained dependencies inherent in physiological and behavioral markers. This represents a significant gap, as recent Transformer architectures like PatchTST [34] are well-suited to address the specific challenges of mobile sensing, such as sensor noise and unavailability[1, 11]. By processing sensor channels independently, these models prevent noise from unstable sensors (e.g., GPS drift) from corrupting the representations of robust signals, establishing a resilient backbone for encoding high-dimensional, multivariate temporal context.

To operationalize this for JITAIs, we couple this encoder with a deep survival analysis decoder. Prior research in large-scale notification management has demonstrated that survival analysis provides a superior fit for engagement modeling by explicitly accounting for the 'right-censored' nature of user interaction, where a non-response is treated as a delay rather than a rejection [51, 53]. Building on this, we adopt a DeepHit-style [21, 22] objective to estimate the full probability mass function (PMF) of response times. This approach avoids the rigid assumptions regarding the relationship between covariates and risk inherent in parametric models (e.g., Weibull), instead learning the joint distribution directly which has been in the past shown to effectively capture complex, non-linear dependencies. [22].

Guided by this architecture, this granular estimation of the survival distribution allows us to operationalize receptivity as a continuous, dynamic state rather than a static label. The design of this pipeline is motivated by the specific structural constraints of mHealth receptivity: a noisy-negative target that demands distributional modeling, heterogeneous sensor streams that require channel-isolated encoding, optimization instability on behavioral time-series, and label scarcity that necessitates self-supervised representation learning. To validate the utility of this approach in real-world mHealth deployments, we target the following research questions:

- **RQ1: Feasibility of Continuous Receptivity Architectures.** Is it feasible to model receptivity as a probabilistic time-to-event distribution using high-dimensional, multivariate mobile sensing data? We investigate the architectural coupling of Channel-Independent Transformers (PatchTST) with survival analysis to capture the nuances of user receptivity.
- **RQ2: Robustness and Stability.** Does the proposed time-to-event framework provide a stable and consistent representation of receptivity across multiple decision horizons? We examine whether a single trained model remains discriminative across cutoffs and whether it captures temporal structure that binary operationalizations fail to represent.
- **RQ3: Strategies for Data Efficiency.** To what extent can self-supervised learning strategies leverage unlabeled sensor traces to mitigate the 'cold-start' problem? We explore whether pre-training via Masked Patch Reconstruction on auxiliary datasets allows the model to learn transferable behavioral representations, thereby accelerating convergence and reducing the dependency on scarce labeled interaction data.

In addressing these research questions, this work makes the following **contributions**:

- We reformulate receptivity as a time-to-event distribution; moving beyond fixed-window binary operationalizations and model receptivity as a probabilistic distribution over response latency, addressing the noisy-negative problem that conflates delayed engagement with non-engagement in prior work.
- We introduce PRISM, a co-designed modeling framework that integrates a channel-independent Transformer encoder, a discrete-time survival objective, and self-supervised pretraining to address key structural constraints of mHealth sensing data, including sensor heterogeneity, optimization instability, and label scarcity.
- Through multi-horizon evaluation, we show that a single trained model remains stable across multiple decision cutoffs, in contrast to binary baselines that require retraining at each threshold or degrade as the response horizon widens. Thus highlighting distributional validity.
- We provide preliminary evidence of generalization by running the formulation on an independent dataset with variable prompt timing. Our findings show consistent trends and suggest that the approach is not tied to fixed-schedule deployments.
- We present the implications for the future of intervention design and scheduling. We show how distributional predictions decouple model training from intervention policies, enabling reasoning about *when* to intervene rather than only *whether* to intervene within a fixed window.

2 BACKGROUND AND RELATED WORK

2.1 State of Receptivity Detection

Early research characterized receptivity as a function of contextual metadata, identifying strong associations with temporal patterns [2, 26, 38, 41], location and connectivity [27, 38, 40, 44], as well as device status and communication logs [10, 40]. These foundational studies extended to inferential traits, linking receptivity to physical activity [35], personality [28], and mental states [44].

Building on these associations, the field has advanced toward real-time predictive modeling. Approaches focusing on cognitive breakpoints [36] and inferred states of boredom [41] demonstrated that timing notifications to natural breaks could significantly reduce response latency. While initial deployments utilizing simple classifiers (e.g., Naive Bayes) yielded mixed results regarding user engagement [31], subsequent iterations have proven more robust. Notably, work by Kunzler et al. [19] and Mishra et al. [30] established that context-aware delivery can improve receptivity rates by over 40% compared to random timing, validating the utility of machine learning in dynamic intervention scheduling.

As predictive capabilities mature, recent works have expanded to scrutinize the systemic implications of automated triggering. King et al. demonstrated that optimizing for receptivity effectively alters the sampling mechanism itself, introducing potential downstream biases in the collected data [17]. Concurrently, Sahu et al. have moved toward a more granular taxonomy, separating the initial *acceptance* of a prompt from the situational *feasibility* of completion [42]. These findings suggest a critical maturation in the field, where, rather than merely predicting engagement, researchers are now 'debugging' the intervention loop, striving to uncover the latent behavioral dynamics that drive these complex interactions.

2.2 Limitations of Binary Operationalization in Receptivity Modeling

The theoretical definition of receptivity emphasizes a nuanced and context-aware state of willingness [33], the prevailing operationalization in mobile health (mHealth) is based heavily on binary classifications. Standard practice designates a user as "receptive" ($y = 1$) if they respond to a JITAI within a fixed time window, most commonly 10 minutes and "non-receptive" ($y = 0$) otherwise [17, 29]. We contend, however, that this binary perspective overlooks the nuanced temporal dynamics of receptivity, reducing a continuous range of engagement to a simple either-or categorization that may constrain the accuracy of mHealth interventions.

Reliance on a fixed threshold introduces methodological challenges, mainly because the duration lacks unique predictive validity. Evidence from the literature demonstrates that while models trained on 10-minute windows outperform random baselines, so do models trained on 1-minute or 5-minute windows [29]; no single window has unique predictive validity, and the choice trades off impulse engagement against deferred availability. Consequently, the selection of the window often involves balancing theoretical intent with practical constraints. To combat the severe class imbalance inherent in the sparse mHealth data, researchers often expand the receptivity window to capture more positive instances (pseudolabeling), effectively adjusting the ground truth to accommodate data sparsity [17]. This flexibility leads to varying optimization targets: a model trained on a narrow window (< 5 minutes) optimizes for "impulse" engagement, detecting immediate interruptibility, whereas a wide window (> 60 minutes) optimizes for "deferred" planning. By standardizing on a middle ground, the community risks optimizing for a compromise that serves neither urgent context-aware prompts nor high-burden therapeutic interventions effectively.

Furthermore, the binary label creates a "noisy negative" by aggregating distinct behavioral states. In a binary framework, a non-response ($y = 0$) does not distinguish between *non-perception* (e.g., the user is physically incapacitated or separated from the device), *active rejection* (the user perceives but dismisses the prompt), and *delayed receptivity* (the user responds after the cutoff). For example, a user responding at 10 minutes and 30 seconds exhibits a fundamentally different cognitive state than one who never engages, yet binary classifiers

treat these as identical negative instances. This over-simplification obscures the mechanism of action, making it difficult for models to learn the nuances between immediate interruptibility and delayed availability.

Finally, this rigid temporal operationalization renders the relationship between intervention complexity and response latency opaque. While the field increasingly recognizes the importance of the *Burden of Treatment* [7], a fixed window prevents the analysis of whether high-burden tasks naturally require longer receptivity horizons than low-burden micro-interactions. By treating all non-responses within the fixed window as uniform failures, the current paradigm structurally excludes the concept of "deferred availability." To advance the field, we propose moving beyond these static thresholds toward a framework that models receptivity as a time-to-event probability distribution. This shift provides the necessary observational resolution to eventually study these dynamics, preserving the fidelity of the urgency signal that binary labels discard.

2.3 Survival Analysis in HCI/mHealth

Survival analysis, historically rooted in medical and epidemiological research, encompasses a set of statistical approaches for analyzing the expected duration of time until one or more events of interest occur, such as the onset of a disease or mechanical failure [53]. Unlike standard regression or classification tasks that predict a snapshot value or label, survival analysis models the underlying probability distribution of the time-to-event.

A defining characteristic of this domain is its ability to handle *censoring*, a phenomenon in which the event of interest is not observed for some subjects within the study window. Survival time generally consists of two components: a beginning point and an endpoint [53]. However, if the event does not occur by the end of the follow-up time, the observation is considered censored. These data points are not "missing" in the traditional sense; they carry valuable information that indicates that the subject "survived" (did not experience the event) for at least time t. Standard machine learning approaches that drop these instances introduce significant bias [3, 14], whereas survival objectives explicitly account for them by maximizing the probability of survival beyond the censoring time.

While traditionally applied to biological lifespans, survival analysis has seen increasing utility in modeling social behavior and human-computer interaction. It has been successfully employed to predict dwell time on web services and model information cascades in social networks [24, 52, 53]. More recently, this approach has been adapted for mobile notification management. Previous works have utilized survival analysis to model the time-to-visit (opening an app) after a notification. For example, Yuan et al. introduced the use of Accelerated Failure-Time (AFT) models assuming a Weibull distribution to predict user engagement [53]. Similarly, Yang et al. extended this via deep learning to optimize notification delivery times [51]. These works demonstrate that modeling the *time* until interaction offers a more granular understanding of user behavior than binary engagement metrics alone.

We propose that survival analysis offers an alternative framework that better aligns with theoretical conceptualizations of receptivity [33]. By modeling the probability distribution over time-to-event rather than a static cutoff, this approach addresses the structural limitations identified in Section 2.2. It preserves temporal granularity, distinguishing between immediate "impulse" engagement and deferred responses without enforcing a compromise threshold. Furthermore, it mitigates the issue of "noisy negatives" by formalizing non-responses as right-censored data, ensuring that unobserved interactions are treated as information that the response time exceeds the observation window, rather than as definitive rejections.

2.4 Deep Learning Approaches to Time-to-Event Analysis

While traditional statistical methods like the Cox Proportional Hazards model have laid the groundwork for survival analysis, they are strictly limited by linearity assumptions and struggle to model the complex, time-varying dependencies inherent in high-dimensional mobile sensing data. To address this, deep learning architectures

evolved from simple Multi-Layer Perceptrons to dynamic frameworks like **DeepHit** [22], which utilizes a multi-task network to learn the joint distribution of survival times and competing events directly, extending to longitudinal settings via Recurrent Neural Networks (RNNs) [21]. However, while RNNs and LSTMs have historically been the standard for sequential sensing tasks, they face significant bottlenecks in capturing long-range dependencies over the extended historical windows necessary to model human routine. Consequently, the field is shifting toward Transformer architectures (e.g., SurvFormer [23]), which utilize self-attention to process sequences in parallel; yet, applying standard Transformers to physiological time-series remains computationally expensive and semantically weak due to point-wise attention. Recent advancements such as the **PatchTST** [34] architecture resolve these constraints by segmenting time-series into sub-series 'patches'. This patching design allows the model to attend to much longer historical contexts along with perserving the local semantic integrity of physical sensor data while enforcing channel independence to prevent noise propagation between heterogeneous sensors.

2.5 Summary: Bridging the Gap

Current approaches to receptivity modeling are constrained by two fundamental limitations: the structural simplification of receptivity into a binary classification problem and the inability of existing architectures to efficiently model long-term behavioral dependencies. First, as detailed in Section 2.2, binary labels create 'noisy negatives' by aggregating distinct states such as physical incapacitation versus active rejection—into a single non-receptive class, thereby obscuring the true mechanisms of user availability. Second, while deep learning has improved predictive power, standard architectures struggle to leverage the vast amounts of unlabeled sensor data available in mHealth. The PRISM framework addresses these gaps by operationalizing receptivity as a **time-to-event probability distribution**, distinguishing between immediate and deferred availability. By integrating a **PatchTST** backbone with a **DeepHit** survival decoder, PRISM captures long-range temporal dependencies while preventing noise propagation via channel independence. Furthermore, PRISM employs a **Self-Supervised Learning (SSL)** strategy via masked patch reconstruction, allowing the model to learn robust behavioral representations from unlabeled data, significantly mitigating the 'cold-start' problem inherent to personalized intervention systems. Together, these design choices address the specific structural constraints of receptivity modeling: noisy-negative targets, heterogeneous sensor streams, and label scarcity.

3 THE PRISM FRAMEWORK

In this section we introduce `PRISM`, a framework designed to model receptivity as a time-to-event probability distribution. PRISM comprises three components: (1) a channel-independent transformer backbone, (2) a probabilistic survival decoder, and (3) a two-stage optimization strategy employing self-supervised learning.

3.1 Context Encoding via Self-Supervised Transformers

PRISM encodes longitudinal behavioral patterns from noisy, unlabeled data using a Channel-Independent PatchTST encoder [34]. We selected this architecture to address the non-stationarity and computational constraints inherent in mobile sensing data.

3.1.1 Reversible Instance Normalization (RevIN). Longitudinal user data often exhibits significant distribution shifts over time (e.g., activity levels changing seasonally). To mitigate this non-stationarity, we employ Reversible Instance Normalization (RevIN) [16]. Prior to encoding, we normalize each instance $\mathbf{x}^{(i)}$ by calculating its mean μ_i and standard deviation σ_i:

$$\hat{\mathbf{x}}^{(i)} = \gamma \left(\frac{\mathbf{x}^{(i)} - \mu_i}{\sigma_i + \epsilon} \right) + \beta \tag{1}$$

where γ and β are learnable affine parameters. This normalization stabilizes the distribution for the encoder, and the inverse operation is applied later to restore the original scale for valid reconstruction loss calculation.

3.1.2 Channel-Independent Patching for Robustness and Efficiency. To process long history windows efficiently, we employ a patching operation $\mathcal{P}$. Given a univariate sensor stream $\mathbf{x}^{(i)}$ of length L, we unfold it into a sequence of N patches with length P and stride S:

$$\mathbf{x}_p^{(i)} = \text{Unfold}(\mathbf{x}^{(i)}) \in \mathbb{R}^{P \times N}, \quad \text{where } N = \left\lfloor \frac{L-P}{S} \right\rfloor + 2 \tag{2}$$

This transformation reduces the memory complexity of the attention mechanism from $O(L^2)$ to $O(N^2) \approx O((L/S)^2)$, enabling PRISM to efficiently process extended historical windows (e.g., 14 days). Capturing long-range dependencies are highly applicable in the mHealth to model multi-scale temporal dynamics, such as circadian rhythms and weekly behavioral motifs, which are often lost in shorter context windows.

Furthermore, mobile sensing data is inherently heterogeneous; a 'bursty' event-based signal (e.g., Screen Unlocks) has a different statistical profile than a continuous state signal (e.g., Battery Level). Standard transformers that mix channels risk allowing high-frequency noise from unstable sensors to corrupt the representations of stable ones. To mitigate this, PRISM enforces *Channel Independence* (CI). Each sensor stream is embedded and processed by the transformer backbone in isolation, sharing weights to learn universal temporal patterns while preventing noise propagation across modalities [34]. In this architecture, patching serves a dual purpose: it acts as a local filter to smooth out sensor jitter, and it quadratically reduces computational complexity, enabling the model to learn representations from high-resolution data without the memory bottleneck of point-wise attention. This design choice is specifically motivated by the structure of mobile sensing data, where the statistical profile of each sensor stream differs (event-based vs. continuous), and where sensor availability varies across participants and devices.

3.2 Probabilistic Receptivity Inference

We frame receptivity as a discrete-time survival analysis problem. The model estimates the probability mass function (PMF) of the "time-to-response" over K discretized time bins.

To balance calibration (accurate probability estimates) with discrimination (correct ranking of risks), we adapt the DeepHit objective [22]. However, the standard ranking loss in DeepHit is prone to gradient explosion during early training dynamics. We introduce a numerical stability control, defining the PRISM objective as:

$$\mathcal{L}_{\text{PRISM}} = \mathcal{L}_{\text{NLL}} + \beta \sum_{i \neq j} A_{ij} \cdot \eta(\hat{F}_i, \hat{F}_j) \tag{3}$$

Here, $\mathcal{L}_{\text{NLL}}$ is the negative log-likelihood of the joint distribution of the first hitting time and event. The second term is a ranking loss weighted by β. The indicator matrix $A_{ij} = \mathbb{I}(y_i = 1, t_i < t_j)$ identifies valid pairs for comparison within the training batch. Intuitively, if instance i is observed to respond at time t_i and instance j responds later (or hasn't responded yet) at t_j, the model should assign a higher cumulative probability of response to i than to j at time t_i. This loss does not compare "responders" vs. "non-responders" as binary classes; rather, it exploits the relative temporal ordering (deltas) between instances to refine the shape of the predicted survival curves, ensuring the model correctly differentiates between immediate and delayed engagement. The two terms play complementary roles: the NLL term calibrates the marginal PMF over time bins, while the ranking term enforces inter-user ordering. In preliminary experiments, optimizing NLL alone produced models that collapsed to the population-average response curve with near-chance inter-user discrimination, whereas optimizing ranking alone was numerically unstable; both terms are required.

The term $\eta(\cdot)$ represents our "Safe" ranking penalty. We modify the standard exponential penalty with a clamping operation C to prevent divergence:

$$\eta(u, v) = \exp\left(C\left(\frac{-(u-v)}{\sigma}, \max = 10.0\right)\right) \tag{4}$$

where $\hat{F}$ is the estimated cumulative distribution function and σ controls the steepness of the penalty. This modification ensures the model navigates the initial optimization landscape without collapsing. This stabilization was necessary because, unlike clinical datasets where event times are well-separated (e.g., months between diagnoses), behavioral time-series contain events at high temporal resolution (minute-level), producing many closely spaced pairwise comparisons that amplify the exponential penalty in early training iterations. The clamp threshold is set so the bounded penalty stays within the `float32` dynamic range while still imposing a gradient large enough to drive optimization. We arrived at this bound after gradient clipping alone (`clip_grad_norm_` at 1.0) failed to prevent NaN recurrence: in our preliminary supervised runs the unbounded ranking term diverged within the first ten epochs, whereas every supervised and pre-trained run conducted with the clamped variant (N=5 folds × multiple training configurations reported in this paper) completed without NaN events.

3.3 Training Process

Our training pipeline consists of two phases designed to leverage unlabelled data while preventing catastrophic forgetting during fine-tuning.

3.3.1 Phase 1: Self-Supervised Pre-training. We first train the backbone on a corpus of unlabelled sensor data using Masked Patch Modeling. We randomly mask part of the input patches and minimize the Mean Squared Error (MSE) of the reconstruction. This forces the model to learn the latent 'grammar' of human routine independent of receptivity labels.

3.3.2 Phase 2: Discriminative Fine-Tuning. We subsequently attach the survival decoder and fine-tune the model. To preserve the temporal features learned during pre-training, we employ Layer-wise Learning Rate Decay (LLRD) [18]. The learning rate η_l for layer l is defined as:

$$\eta_l = \eta_{\text{head}} \cdot \lambda^{D-l} \tag{5}$$

where η_{head} is the learning rate for the survival head, D is the network depth, and $\lambda < 1$ is the decay factor. This differential optimization acts as a "gradient damper" for the lower layers, ensuring the backbone retains its generalizability while the head rapidly adapts to the receptivity task. We arrived at LLRD after observing two failure modes in preliminary experiments: uniform fine-tuning allowed the randomly initialized head to shatter the pre-trained backbone features (catastrophic forgetting), while hard-freezing the backbone prevented adaptation on participants whose behavior differed from the pre-training distribution (negative transfer). The differential learning rates resolve both by keeping the backbone stable enough to preserve pre-trained structure while allowing the head to adapt rapidly.

4 DATASET

4.1 Dataset Used

The data utilized in this work originate from three distinct deployment phases of 'LvL UP', a smartphone-based holistic lifestyle coaching intervention developed to improve health behaviors and mental well-being [4]. The intervention integrates automated conversational agent–led health coaching sessions with complementary self-regulatory tools (activity tracker, food diary, journal, and slow-paced breathing) [45]. All three studies were conducted in Singapore. We thank the authors of the LvL UP trials for making their data available to us for

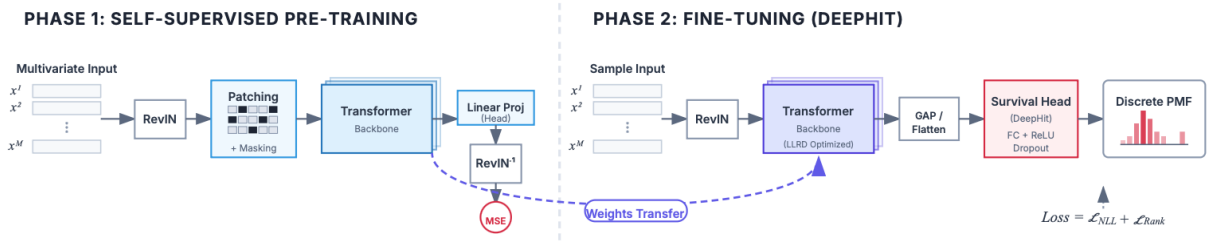


Fig. 1. **The PRISM Framework.** The architecture utilizes a two-stage training protocol. **Phase 1 (Left):** The model engages in Self-Supervised Learning on unlabeled sensor data, optimizing a reconstruction objective (MSE) to learn temporal dependencies via random masking. **Phase 2 (Right):** The pre-trained backbone weights are transferred to the downstream task. A DeepHit Survival Head is attached to predict receptivity, and the network is fine-tuned using Layer-wise Learning Rate Decay (LLRD) to preserve the feature representations learned in Phase 1.

this analysis. To capture the context of user engagement, the application includes an integrated passive sensing module that runs continuously in the background.

To address the scarcity of labeled receptivity data, we adopt a split-source strategy. We aggregate the longitudinal sensor streams from the two preliminary studies (Feasibility and Pilot) to form a large-scale unlabeled corpus for self-supervised pre-training to learn robust representations of habitual behavior. Data from the Main Trial is reserved exclusively for supervised fine-tuning and evaluation, providing the ground-truth receptivity labels required for the downstream survival analysis task. The specific characteristics of the three deployment cohorts are as follows:

- **LvL UP Feasibility Study [25]:** A 21-day 'in-the-wild' trial with Singaporean adults ($n = 99$) to assess the technical feasibility, app engagement, and user satisfaction of the intervention.
- **LvL UP Pilot Trial [54]:** An 8-week sequential, multiple assignment, randomized controlled pilot study ($n = 123$) aimed to evaluate the feasibility and preliminary efficacy of the intervention in adults at risk of developing non-communicable diseases or common mental disorders.
- **LvL UP Main Trial [5]:** A 12-month sequential, multiple assignment, randomized controlled trial ($n = 678$) aimed to evaluate the effectiveness of the intervention in adults at risk of developing non-communicable diseases or common mental health disorders.

4.2 Intervention Protocol and Sampling Strategy

Our analysis uses an app use reminder notification, which was delivered by the LvL UP app daily at 09:00 local time. This prompt serves as a consistent behavioral anchor, encouraging users to engage with the application's lifestyle coaching tools.

While receptivity research often relies on random time-stratified sampling (e.g., Experience Sampling) to capture temporal dynamics, this intervention adheres to a fixed daily schedule. We intentionally adopt this design to control for circadian and time-of-day effects, thereby isolating day-to-day variation in contextual receptivity. This controlled setting provides a conservative and internally valid testbed for evaluating our central contribution: operationalizing receptivity as a continuous time-to-event process rather than a fixed-window binary outcome. Under this protocol, variability in response latency arises primarily from changing environmental and behavioral contexts (e.g., location, recent activity, sleep history), rather than from differences in delivery time. Consequently, the model must learn to infer receptivity from sensed context rather than relying on simple temporal heuristics. We define our target variable as the latency between this 09:00 h prompt and the user's subsequent interaction with the app. To ensure a clean signal, we excluded days where participants received

Table 1. Summary of passive sensing features and data types.

Category	Continuous Features	Categorical Features
Device Status	`battery_level`, `screen_flips`	`batt_status`, `batt_state`, `scr_state`
Notifications	`notif_rem_unique`, `notif_rem_tot`	*None*
Mobility	`steps`, `loc_speed`, `loc_acc`	`loc_context`, `weekday`
Info Flags	*None*	`batt/notif/scr/step/loc_flags`

additional, non-routine notifications (e.g., ad-hoc reminders) to eliminate confounding effects. Despite the fixed delivery time, the longitudinal nature of the Main Trial provides substantial within-user and between-user variation in context and engagement patterns, enabling us to examine how receptivity evolves over time

4.3 Feature Extraction and Processing

We processed the raw Android sensor streams into a unified time-series dataset with a 1-minute resolution. As mobile sensing data is inherently sparse and event-driven (recording values only upon state changes), we aligned all sensors to a continuous temporal grid using a 'sample-and-hold' strategy. We forward-filled values for state-based features (Battery, Screen, Location) up to a sensor-specific horizon determined by the 99th percentile of inter-arrival times. We explicitly marked gaps exceeding this horizon as missing. For accumulative features (e.g., Steps), we uniformly distributed aggregate counts across the reporting intervals.

To preserve data integrity for downstream sequence modeling, we generated a metadata channel (`data_info_flag`) indicating whether each timestep represented a raw observation, an imputed value, or a missing period. This creates a dense representation suitable for the Transformer encoder while retaining the semantic distinction between "no data" and "no change." Finally, we normalized continuous features and encoded categorical features for embedding layers. We included an additional feature representing the day of intervention to account for the temporal evolution of user engagement. Table 1 lists the features used for the analysis and model training.

4.4 Sample Definition and Cohorts

We defined a sample as a multi-feature array representing a 24-hour context window, comprising 1,440 minute-level time steps. To address the scarcity of labeled receptivity data, we adopted a split-source strategy:

- **Labeled Dataset (Evaluation):** Derived from the LvL UP Main Trial. From the 678 enrolled participants, we retained those assigned to the intervention arm and whose passive sensing was collected through the Android pipeline described above. This selection yielded 3,462 samples from 101 participants, with individual contributions ranging from 3 to 96 samples. We used this strictly for supervised fine-tuning and evaluation.
- **Unlabeled Dataset (Pre-training):** We used pooled data from the two preliminary studies (Feasibility and Pilot) to create a large-scale corpus for self-supervised learning(SSL). We initially identified 2,447 potential samples but applied a quality filter retaining only samples with at least 50% valid data (defined as time-steps which are not flagged as 'no data' via the metadata channel described in Section 4.3). This criterion excluded 1,531 sparse instances, resulting in 916 samples for the self-training phase. We enforced this quality threshold exclusively for the self-supervised phase. Prior work emphasizes that robust representation learning requires sufficient information density to solve the reconstruction objective; if the input is predominantly missing, the effectiveness of the masking task degrades [20, 34]. To ensure the model learns from valid behavioral signals rather than fitting to noise or missingness artifacts, we prioritized high-quality dense samples to train the encoder. In contrast, labeled receptivity data is scarce. We therefore avoided filtering the evaluation set to maximize the utilization of ground-truth supervision and to ensure our evaluation reflects the realistic

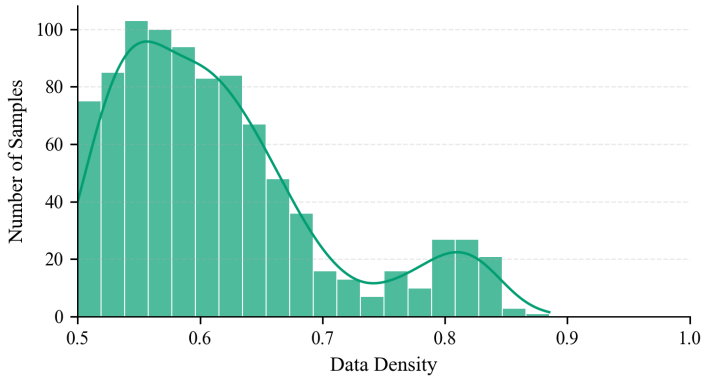


Fig. 2. Distribution of retained samples for self-supervised learning (density ≥ 0.5).

sparsity of in-the-wild deployments, relying on the pre-trained features to handle sparse inputs during fine-tuning [18]. Figure 2 illustrates the distribution of these retained samples.

5 METHODOLOGY

5.1 Experimental Setup

We evaluated the PRISM framework using a 24-hour historical context window ($L = 1440$ minutes) to predict user receptivity. We employed a 5-fold Group K-Cross Validation strategy to ensure robustness. All data from a single participant remained strictly within a single fold (either train or test) to prevent data leakage and to measure generalization to unseen users. We maintained a consistent encoder-only backbone across all models ($L = 1440$, patch length $P = 16$, stride $S = 8$). The transformer depth consisted of 3 layers with a model dimension of $D = 64$ and 4 attention heads. We optimized the objective defined in Section 3.2, adapting the DeepHit loss with a ranking penalty for numerical stability ($\alpha = 0.5$).

5.2 Target Definition and Evaluation Metrics

We framed the prediction task as discrete-time survival analysis with administrative truncation. We defined an actionable horizon of 180 minutes post-notification, discretizing this interval $[0, 180]$ into $K = 24$ bins. Unlike clinical trials with random censoring, we assume an eventual response; we mapped valid responses to their respective bins and assigned outliers (responses > 180 minutes or missing) to the final time bin (K). This approach avoids unbounded regression targets while categorizing late responses as effectively 'non-receptive.'

Since PRISM outputs a probability distribution (Time-to-Event) while baselines output binary predictions, we projected the probabilistic output into the binary classification space for fair comparison. For a given instance i, PRISM predicts the cumulative distribution function $\hat{F}_i(t) = P(T \leq t)$. We extracted the predicted probability of response within the target horizon $\tau = 15$ minutes:

$$\hat{y}_{prism} = \hat{F}_i(\tau) \tag{6}$$

We treated this value as the positive class probability, allowing us to compute standard metrics (AUC, F1-score) across all models regardless of their original output format.

5.3 Baseline Assessments

We evaluated PRISM under two complementary assessments. The first matches the prevailing operationalization of receptivity in JITAI systems by fixing a single decision cutoff at $\tau = 15$ min and comparing all models as binary classifiers at that cutoff. This projection allows direct comparison with prior receptivity models, which are typically evaluated under fixed-window binary formulations. The second exercises PRISM's distributional output by reading it at multiple cutoffs, with baselines either retrained per cutoff or scored at every cutoff from a single fit. We describe each assessment and its baselines below. All baselines used the same 5-fold group-partitioned split over participants as PRISM, so per-fold results are directly comparable.

5.3.1 Comparative Performance Analysis. The first assessment fixes $\tau = 15$ min, a cutoff within the short fixed-window range used in prior receptivity work [17, 29, 40] and aligned to a bin edge of PRISM's 7.5-minute discretization grid. At this cutoff every model emits one positive-class score per sample, and we computed a single AUC per model per fold. PRISM's distributional output is collapsed to a scalar via the projection rule defined in Section 5.2, $\hat{y} = \hat{F}(\tau)$. This setting tests whether PRISM's time-to-event formulation can match conventional binary classifiers for the task of receptivity. We compared PRISM in this setting against two binary baselines, described below: a tabular gradient-boosted classifier (**XGBoost (Binary)**) and a recurrent deep-sequence classifier (**LSTM (Binary)**).

XGBoost (Binary). A gradient-boosted decision tree trained on static "snapshot" features, following the baseline established in prior JITAI receptivity studies [29]. The model received a tabular vector of the last known values for all context features (e.g., battery level, step count, location) at the moment of intervention and produced a binary classification of receptivity at $\tau = 15$ min.

LSTM (Binary). A recurrent deep-sequence baseline trained as a binary classifier at the same $\tau = 15$ min cutoff. LSTMs remain a standard reference for multivariate sequence modeling in mobile and physiological sensing [13, 17], and including a binary-headed sequence model alongside XGBoost helps separate the contribution of the sequence backbone from the choice of tabular features. The model consumed the same $L = 1440$ minute-level context window and input pipeline as PRISM (categorical embeddings, instance-normalized continuous channels). A 2-layer bidirectional LSTM with hidden size 128 produced an output sequence whose mean across time was concatenated with the final time-step representation and passed through a two-layer MLP (128 units, ReLU, dropout 0.1) to a sigmoid-activated scalar head. We trained with binary cross-entropy on the $\tau = 15$ min labels (AdamW, gradient clipping at 1.0, early stopping on validation AUC).[1]

5.3.2 Evaluation Across Response Horizons. PRISM's time-to-event output can be read at any cutoff, not only at $\tau = 15$ min. The second assessment tests whether that flexibility preserves predictive performance: we evaluated each model at cutoffs $\tau \in \{15, 30, 60, 90, 120\}$ min, chosen to fall on bin edges of PRISM's 7.5-minute discretization grid so no interpolation was required, and applied the projection rule defined in Section 5.2 at every τ. A single PRISM fit per fold produces the full distribution and is read at each cutoff without retraining, isolating whether the same trained model carries a usable signal across horizons. In this setting we compared PRISM against two per-horizon binary specialists (**XGBoost (Binary)** and **LSTM (Binary)**) and one one-fit sequence baseline (**LSTM (Regression)**), each described below.

XGBoost (Binary, per-horizon specialist). XGBoost is a binary classifier by construction and cannot project to multiple cutoffs from a single fit. We therefore retrained a separate model at every τ, with labels re-thresholded at that cutoff. Each model is therefore a *specialist*: tuned exclusively for its own horizon rather than expected to

[1]We swept LSTM/GRU cells, 1–3 layers, hidden sizes $\in \{64, 128, 256\}$, uni- and bidirectional variants, and dropout settings; the configuration reported here (2-layer biLSTM, hidden 128) achieved the best validation AUC and is used throughout.

generalize across τ. We instantiate this procedure at every τ we evaluate, giving the classical baseline its strongest available footing at each horizon.

LSTM (Binary, per-horizon specialist). We applied the same per-horizon specialist procedure to the LSTM (Binary) model defined in Section 5.3.1: a separate binary classifier was trained at every τ, with labels re-thresholded at that cutoff and the same backbone, input pipeline, and training configuration retained across horizons. This pairs the tabular specialist with a sequence-model specialist, isolating whether per-horizon retraining of a recurrent backbone closes the gap to PRISM's single-fit distributional output.

LSTM (Regression). To pair PRISM's one-fit property with a sequence-model counterpart, we re-used the LSTM backbone described above with a regression head: a Softplus-activated scalar output emitting a non-negative response latency in minutes, trained with an L_1 loss and the regression target clipped to the 180-minute horizon (AdamW, gradient clipping at 1.0, early stopping on AUC at $\tau = 15$ min). A scalar latency estimate ranks samples on expected response time and therefore yields a valid positive-class score at any τ via sign flipping, which lets a single trained model be scored at every cutoff. A per-cutoff classifier head would instead commit the training objective to one cutoff and conflate modeling capacity with cutoff-specific specialization, which is why we chose the regression head for the horizon-sweep assessment.

5.4 Ablations

We further investigate the architectural contribution by evaluating **PRISM w/o SSL**. This ablation isolates the benefits of the pre-training strategy. The model utilizes the identical backbone but is initialized randomly and trained via supervised learning. This comparison isolates the impact of self-supervised representation learning from the sequence modeling capability of the transformer architecture.

5.5 Training Protocols

5.5.1 Supervised Training Strategy. For the **PRISM w/o SSL** ablation, we trained the model for a maximum of 50 epochs with early stopping (patience of 15 epochs). Optimization used AdamW with a global weight decay of 1e−3. We employed Layer-wise Learning Rate Decay (LLRD) to train the deep transformer on the limited labeled data. We assigned a lower learning rate to the backbone ($\eta_{\text{backbone}} = 2\text{e}{-5}$) to preserve stability, while the survival head (256 units, ReLU, Dropout 0.2) updated rapidly ($\eta_{\text{head}} = 1\text{e}{-3}$). A 'ReduceLROnPlateau' scheduler (factor 0.5, patience 5) managed the learning rate adjustments.

5.5.2 Self-Supervised Pre-training. The full **PRISM** framework incorporates a Self-Supervised Learning (SSL) pre-training phase. This strategy is motivated by a structural asymmetry in mHealth data: passive sensing streams are relatively easier to collect than ground-truth receptivity labels, which require explicit user interaction with intervention prompts. This phase leverages a large-scale unlabeled dataset comprising diverse behavioral patterns from participant history not associated with receptivity labels. During the first phase (Pre-training), the backbone utilized a Masked Auto-Encoding (MAE) objective. We randomly masked 40% of the input patches and minimized the Mean Squared Error (MSE) of the reconstruction. This phase ran for 50 epochs using an AdamW optimizer ($\eta = 1\text{e}{-3}$) coupled with a Cosine Annealing scheduler ($\eta_{\min} = 1\text{e}{-5}$).

In the second phase (Fine-Tuning), we transferred the pre-trained weights to the supervised task. We increased the regularization in the head (Dropout increased to 0.25) to prevent overfitting to the smaller labeled set. The fine-tuning process mirrored the supervised baseline: 50 epochs, early stopping (patience 15), and the LLRD configuration ($\eta_{\text{backbone}} = 2\text{e}{-5}$, $\eta_{\text{head}} = 1\text{e}{-3}$) to adapt the learned representations to the receptivity task.

5.5.3 Linear Probing and Fine-Tuning (LP-FT). We implemented the **LP-FT** (Linear Probing then Fine-Tuning) protocol to address potential distortion of pre-trained features during the initial fine-tuning phase[18]. This

training occurred in two stages. In Stage 1 (Linear Probing), we froze the transformer backbone, allowing gradients to flow only to the randomly initialized survival head, embeddings, and normalization layers. We trained for 20 epochs with a learning rate of 1e−3 for all active parameters. This aligns the head with the pre-trained features before the features shift. In Stage 2 (Full Fine-Tuning), we unfroze the backbone and proceeded with full fine-tuning for an additional 30 epochs. We re-introduced LLRD ($\eta_{\text{backbone}} = 2\text{e}{-5}$, $\eta_{\text{head}} = 1\text{e}{-3}$) to fine-tune the entire network. This protocol serves as an experimental upper bound, trading increased training complexity for stability.

6 RESULTS

6.1 Model Fidelity & Interpretation

We visualize the model's native output in Figure 3 (Left) to demonstrate the granularity of the learned representations. Unlike standard receptivity classifiers that output a single binary decision, PRISM generates a full probability mass function (PMF) over the actionable 180-minute horizon. The y-axis represents the probability of a receptive state occurring within a specific time bin, following a notification. In this plot, we average the distributions across the Early, Mid, and Late phases of the intervention for a single participant to illustrate longitudinal behavioral shifts.

The distinct morphologies of these curves highlight the model's ability to capture latent dynamics. For this participant, the Mid phase (Day 22-37) is characterized by a sharp peak within the first 30 minutes, indicating a higher probability of rapid engagement compared to other phases. Conversely, the Late phase exhibits a pronounced density increase in the final time bin (160–180 mins). This divergence in the model's learned representation provides a quantitative lens to investigate phenomena such as 'prompt fatigue,' suggesting a state where the user's intent to respond persists but latency increases. Rather than treating these late responses as simple failures, the model learns to adjust the shape of urgency based on the intervention phase, allowing for a deeper inspection of user compliance that binary metrics obscure.

Figure 4 (Right) illustrates the training dynamics, providing insight into the stability offered by Self-Supervised Learning (SSL). We report two metrics: *Ranking Loss*, which measures the model's ability to correctly order risk times (e.g., ensuring a 10-minute response is ranked higher than a 60-minute response), and *Entropy*, which quantifies the uncertainty of the predicted distribution. The shaded regions denote the standard deviation across the 5 cross-validation folds.

First, regarding **Entropy**, PRISM (initialized with SSL) starts with higher entropy compared to the version trained from scratch (PRISM w/o SSL). This indicates that the pre-trained backbone initializes the model with a broader prior, allowing it to explore the probability space before narrowing down to a confident prediction. In terms of **Ranking Loss**, the SSL-initialized model demonstrates a smoother descent and tighter error bounds (smaller shaded area) compared to the baseline. This suggests that while both models learn to discriminate, the self-supervised initialization provides a more stable optimization landscape, reducing variance across different data folds and preventing the model from collapsing into local minima.

Collectively, these qualitative and quantitative insights answer **RQ1**. By architecting a framework that couples a Channel-Independent Transformer with a probabilistic survival decoder, we demonstrate the feasibility of modeling receptivity not as a static binary outcome, but as a time-to-event distribution.

6.2 Comparative Performance Analysis

We compare **PRISM** against a static tabular baseline (XGBoost) and a recurrent deep-sequence binary classifier (LSTM (Binary); Section 5.3.1), both representative of the standard binary operationalization used in prior receptivity work. We additionally report an ablation that removes pre-training (PRISM w/o SSL) and a transfer variant we also examined (PRISM w/ LP-FT). Table 2 summarizes the results.

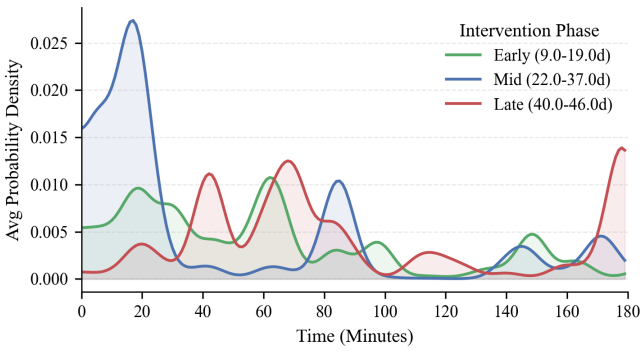


Fig. 3. Single-participant receptivity forecast (24 hours of historic time steps as input). The model predicts a density over the subsequent 180-minute window, averaged across all instances of the participant for a given intervention phase.

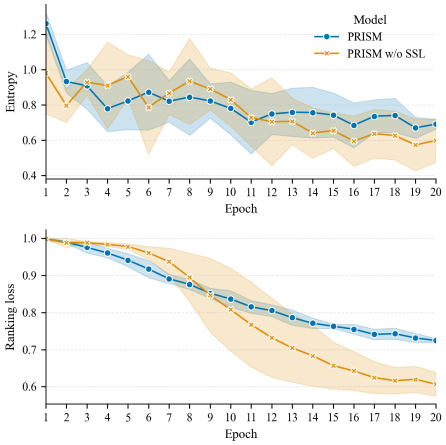


Fig. 4. Training dynamics comparison. Metrics shown are Ranking Loss (Bottom) and Entropy (Top) with shaded regions indicating standard deviation across folds.

This setting is meant as a parity check against the historic benchmark in the receptivity literature, where the task is fixed at a single short cutoff. **XGBoost**, **LSTM (Binary)**, and the supervised **PRISM w/o SSL** ablation cluster within a narrow band at $\tau = 15$ min, with PRISM w/o SSL leading the group only marginally. The two binary baselines optimize directly against the binary $\tau = 15$ min label; PRISM w/o SSL is trained to model the full time-to-event distribution and is collapsed to a scalar score only at evaluation time. The survival-based formulation remains competitive with the binary baselines on their own task without any pre-training, while additionally producing a full response-time distribution that the baselines cannot recover; the richer output therefore comes alongside, rather than at the cost of, parity at the standard cutoff.

The full **PRISM** model, initialized via self-supervised pre-training on unlabeled sensor traces, lifts median AUC well above this cluster. Figure 4 suggests where the gain comes from: PRISM w/o SSL converges to lower training entropy and ranking-loss values, but this faster convergence does not translate to better held-out performance. The SSL-initialized model retains higher entropy throughout training and reaches a higher validation AUC, consistent with pre-training acting as a regularizer that supplies stable representations of habitual behavior from the unlabeled traces. Further tuning of the transfer step appears to leave additional headroom: the **PRISM w/ LP-FT** variant, which aligns the survival head against a frozen pre-trained backbone before unfreezing it, posts a higher median AUC and a tighter dispersion across folds.

These results speak to the parity portion of **RQ2**: under the conventional $\tau = 15$ min operationalization, the time-to-event framework holds parity with the binary baselines without pre-training and surpasses them once SSL is introduced, with fold-level dispersion comparable to the tabular baseline and tighter than the deep-sequence ones (Table 2).

6.3 Evaluation Across Response Horizons

We evaluated all four models across the deployment horizons defined in Section 5.3.2. Figure 5 reports mean binary AUC with standard deviation across the 5 group-partitioned folds.

Table 2. Performance benchmark at $\tau = 15$ min: median AUC and dispersion (MAD) across 5-fold cross-validation. Improvement (%) is reported relative to the XGBoost baseline.

Metric	XGBoost	LSTM (Binary)	PRISM w/o SSL	PRISM	PRISM w/ LP-FT
Median AUC	0.6413	0.6462	0.6469	0.7099	0.7405
Dispersion (MAD)	±0.0180	±0.0359	±0.0553	±0.0218	±0.0053
Improvement (%)	–	+0.76%	+0.87%	+10.70%	+15.47%

The four models split along two axes that the figure makes visible. By *training contract*, **XGBoost (Binary)** and **LSTM (Binary)** are per-horizon specialists retrained at each τ, while **PRISM** and **LSTM (Regression)** are one-fit models scored at every τ from a single trained model. By *behaviour across* τ, two of the models decline monotonically as the cutoff widens and two stay comparatively flat. The split by behaviour does not follow the split by contract: one of each pair sits in each group.

Two of the four models trend downward as the cutoff widens. **XGBoost (Binary)** (per-horizon tabular specialist) and **LSTM (Regression)** (one-fit scalar latency) both decline monotonically from $\tau = 15$ to $\tau = 120$, with LSTM (Regression) sitting below the other models throughout. The remaining two, **PRISM** and **LSTM (Binary)**, stay comparatively stable across τ but reach that stability by different means: LSTM (Binary) is a per-horizon specialist, i.e., it needs to be specifically retrained for each τ cutoff, necessitating five separate training runs, while *PRISM produces the same five readings from just a single fit by projecting its predicted CDF.*

The consistency of PRISM across cutoffs also highlights what the distributional output affords beyond binary projection. The predicted distribution encodes expected latency, the spread of plausible response times, and the probability mass beyond any chosen cutoff, all of which are discarded when reduced to a single threshold. While we do not directly exploit these signals in the present binary-AUC evaluation, their stability across τ suggests that the model is capturing a coherent temporal structure of receptivity rather than overfitting to a specific window. Taken together with the result in Section 6.2, this analysis answers **RQ2**: PRISM matches binary baselines under the conventional $\tau = 15$ min operationalization, benefits from self-supervised pre-training, and remains discriminative across cutoffs where the tabular per-horizon specialist loses ground. We return to the implications of this distributional representation in Section 7.

6.4 Label Efficiency Analysis

This section reads the SSL contribution alongside the architectural reading developed in Section 6.3: where the multi-horizon analysis tests what the distributional head delivers from a single fit, this analysis tests what self-supervised pre-training delivers as labeled data is reduced. We conducted a sensitivity analysis by training the models on progressively reduced subsets of the *labeled* participant pool (10%, 30%, 50%, 70%, 80%, 100%). In this setup, the XGBoost baseline is limited strictly to the reduced labeled set, whereas PRISM benefits from the self-supervised reconstruction task on the underlying sensor features before fine-tuning on the available labels. The test set remained fixed across all experimental conditions, serving as a consistent benchmark. Figure 6 compares the median AUC of PRISM against the XGBoost baseline.

We observe that **PRISM** consistently outperforms the strong gradient-boosting baseline across all data fractions. Notably, even with only 10% of the training participants available, the self-supervised model maintains competitive performance, offering a practical solution to the 'cold-start' problem common in mHealth deployments. This resilience highlights the efficacy of the masked reconstruction objective; by forcing the model to infer missing sensor patches from context, pre-training equips the encoder with a generalized understanding of behavioral dynamics that transfers efficiently to the downstream receptivity task.

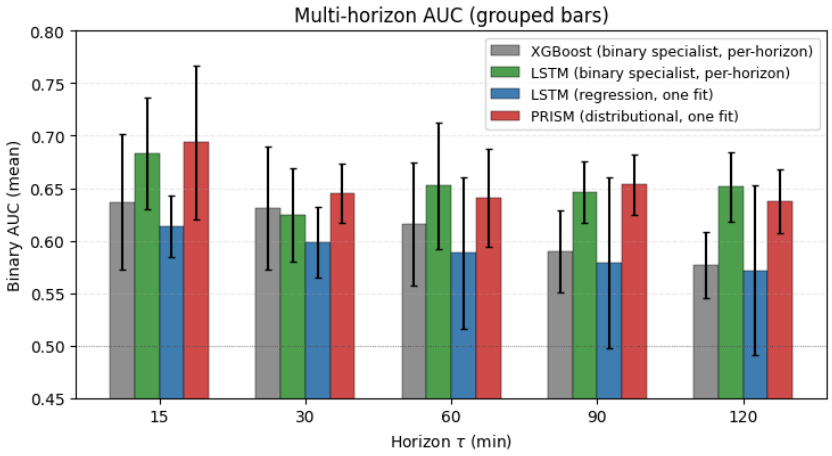


Fig. 5. Mean binary AUC (± standard deviation across 5 `GroupKFold` folds) at horizons $\tau \in \{15, 30, 60, 90, 120\}$ min. **PRISM** (distributional, one fit) is trained once per fold (early-stopped at $\tau = 15$) and evaluated by projecting its predicted CDF at each τ. **XGBoost (Binary)** and **LSTM (Binary)** are per-horizon specialists, retrained at each cutoff. **LSTM (Regression)** (one fit) is scored at every τ by sign-flipped predicted latency.

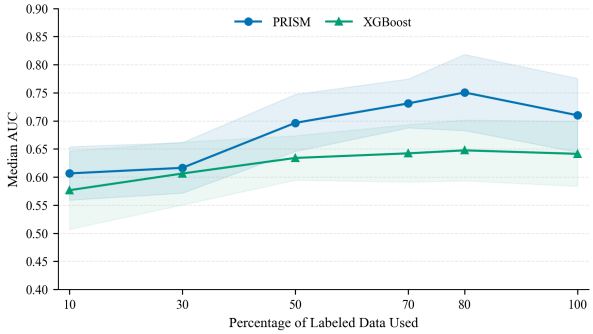


Fig. 6. Model performance (Median AUC) evaluated across varying training set fractions (10%–100%) with a fixed test set, with shaded regions indicating standard deviation across folds

These findings provide a substantive answer to **RQ3**. The model's ability to retain predictive power in data-sparse regimes confirms that learning generalizable representations from passive sensing data is a viable strategy for overcoming the label scarcity bottleneck. By reducing the dependence on extensive participant history, PRISM demonstrates that sophisticated deep learning architectures can be effectively adapted to the constraints of real-world intervention delivery.

6.5 Looking Deeper into PRISM

While the primary research questions (RQ1–RQ3) have been addressed, demonstrating the feasibility, robustness, and efficiency of our approach, we conduct a deeper examination of the PRISM architecture's internal mechanisms.

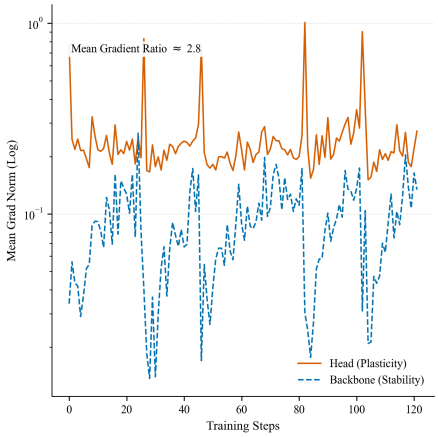


Fig. 7. Differential Learning Dynamics. Gradient logs show the survival head learning ∼30x faster than the backbone, balancing plasticity and stability.

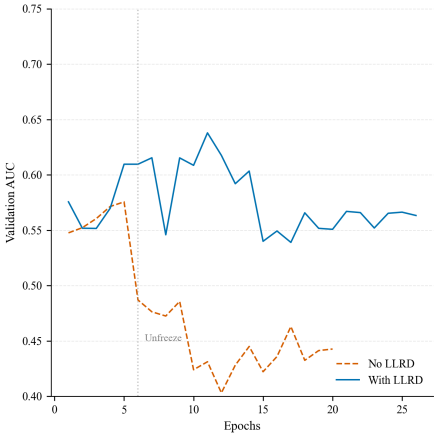


Fig. 8. Impact on Transfer Performance. Validation AUC confirms that LLRD prevents the catastrophic forgetting observed in the naive fine-tuning baseline.

6.5.1 Ablation Study: Impact of Layer-wise Learning Rate Decay. Integrating a pre-trained time-series backbone into a downstream task typically presents a stability-plasticity dilemma. A high learning rate facilitates rapid adaptation of the randomly initialized survival head (plasticity) but risks destroying the pre-trained feature representations (catastrophic forgetting). Conversely, a low learning rate preserves the backbone (stability) but may fail to align the head sufficiently for the receptive task (negative transfer).

We addressed this by implementing Layer-wise Learning Rate Decay (LLRD). In our experiments, we assigned a high learning rate ($\eta_{\text{head}} = 1\text{e}{-3}$) to the survival head and a significantly lower rate ($\eta_{\text{backbone}} = 2\text{e}{-5}$) to the backbone.

Figure 7 traces this differential learning dynamic: head and backbone gradients sit within roughly an order of magnitude of each other, so under AdamW the head-to-backbone per-step update ratio is set primarily by the 50× learning-rate differential, keeping the backbone close to its pre-trained state while the head adapts rapidly. Figure 8 demonstrates the consequence of this mechanism on the validation performance. Without LLRD (dashed red line), the model suffers an immediate performance collapse upon unfreezing the backbone, failing to recover. In contrast, the LLRD-enabled model (solid blue line) transitions smoothly from the frozen phase, continuously improving AUC without degradation.

These results indicate that careful control of gradient flow is critical for preserving the transferable temporal representations learned during self-supervised pre-training. More broadly, they demonstrate that PRISM can be reliably adapted to new receptivity datasets without extensive manual tuning, supporting its practical use in longitudinal deployments where models must be periodically updated as new data become available.

6.5.2 Input Context Window & Feature Reliance. To understand the temporal dependencies of receptivity, we evaluated PRISM on a shorter 3-hour input context window (180 timesteps) versus the standard 24-hour window (1440 timesteps) as reported above. The reduction in context showed similar performance in the receptivity within 15 minutes metric. The 3-hour model achieved an AUC of **0.7159**, comparable to the 24-hour model's **0.7099**. While the predictive performance remained consistent, an analysis of Permutation Feature Importance (PFI) reveals both distinct commonalities and key divergences in the prediction mechanism (Figure 9).

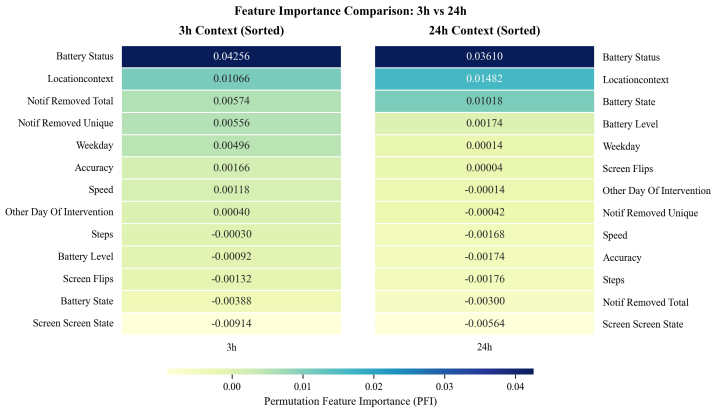


Fig. 9. Permutation Feature Importance (PFI) comparison between 3h and 24h context windows.

Consistent with prior mobile sensing literature [15, 29, 40], **Battery Status** and **Location Context** emerge as dominant predictors across both time horizons (Figure 9). This consistency aligns with the historical view that device state and semantic location are fundamental constraints on receptivity. However, the models diverge in their utilization of high-frequency signals. In the 3-hour model, features related to immediate interaction, such as `notif_data_removed_total`, appear to play a more central role. This sensor tracks the volume of notification dismissals per timestep, effectively serving as a high-resolution proxy for active phone engagement. The elevated importance of such interaction markers in the short-context model highlights an adaptation to *knowledge availability*. By restricting the input window, we effectively 'blind' the model to long-term temporal dependencies, forcing it to compensate by capitalizing on immediate dynamic states. Conversely, the 24-hour model, possessing the broader historical context, can leverage features that capture routine stability. Thus, the PFI shift reflects the model's distinct strategies for extracting signal from the available information horizon.

From a practical perspective, this suggests that the framework offers a configurable trade-off between conceptual depth and computational cost. Rather than enforcing a single rigid architecture, the approach enables designers to align the modeling horizon with their specific system constraints, favoring short-context efficiency or long-context robustness as the deployment environment requires.

6.6 Generalization to Other Studies

The primary evaluation in Sections 6.2–6.3 was conducted on the LvL UP Main Trial, which uses a fixed-time prompt protocol. To probe behavior under a different delivery regime, we conducted a follow-up evaluation on **HAPPY** [43], an independently conducted one-year randomized controlled trial that adapted the LvL UP application and sensing pipeline, but with chatbot-initiated coaching messages delivered at varied times across the day. The HAPPY team made their labeled sensor traces available to us, enabling evaluation on the same sensing and feature pipeline but a different cohort and prompt schedule. We discuss the similarities (sensing pipeline, feature schema, pre-training corpus) and differences (cohort, prompt source, prompt timing) in this section. Given the shared pipeline and differences in deployment conditions, we treat this analysis as a transfer check on generalization rather than a re-engineering of the model.

Some key similarities and differences are:

Table 3. Preliminary transfer evaluation on HAPPY ($N = 865$, 50 participants, variable delivery timing). Median AUC at $\tau = 15$ min and dispersion (MAD) across 5 `GroupKFold` folds; improvement (%) is reported relative to the XGBoost baseline.

Metric	XGBoost	LSTM (Binary)	PRISM	PRISM w/ LP-FT
Median AUC	0.6477	0.6557	**0.6929**	0.6646
Dispersion (MAD)	±0.0082	±0.0483	±0.0442	±0.0506
Improvement (%)	–	+1.24%	+6.98%	+2.61%

- **Application and sensing pipeline (shared).** Both cohorts received the LvL UP application together with the same passive sensing module, so the feature schema in Section 4.3 carried over to HAPPY without modification, and the PRISM input pipeline did not need to be re-engineered.
- **Population (different).** The LvL UP Main Trial recruited Singapore-based adults at risk of non-communicable diseases or common mental disorders. HAPPY recruited post-gestational diabetes women in Singapore.
- **Prompt source and timing (different).** Our primary assessment on LvL UP was anchored to the daily tool-use reminder delivered at 09:00 local time (Section 4), giving a fixed-time anchor across all participants. On HAPPY we used the chatbot-initiated coaching messages: each scripted conversation is opened by an 'initiating message' delivered as a standard mobile notification, and these initiating messages reached participants at varied times across the day. Receptivity in HAPPY is therefore measured against a different message type, on a different schedule.
- **Pre-training corpus (reused from the preliminary trials).** We did not have access to an in-cohort unlabeled corpus for HAPPY, and the smaller participant pool also made it impractical to partition the available data into a separate SSL split. We therefore reused the unlabeled corpus described in Section 4.4 for pre-training, and fine-tuned only on HAPPY labels. This setup turns the comparison into a preliminary check of representation transfer: it asks whether habitual-behavior representations learned by SSL on one population's sensor traces remain useful when fine-tuned against a different population's labels and a different prompt regime on the same sensing pipeline.
- **Device platform.** We restricted the HAPPY analysis to Android participants, matching the Android-only feature set used throughout this work.

The HAPPY analysis cohort yielded $N = 865$ labeled samples from 50 participants. We retained the 24-hour input window, the 5-fold group-partitioned split over participants, and the training protocol described in Section 5.

On HAPPY (Table 3), we observe a similar performance trend as in the LvL UP evaluation (Sections 6.2–6.3). The two binary baselines cluster closely at the 15-minute cutoff (XGBoost 0.6477, LSTM (Binary) 0.6557), while PRISM with SSL pre-training leads with a median AUC of 0.6929, a +6.98% improvement over XGBoost. Although the absolute AUCs are lower across all models, the findings seem consistent with increased variability in delivery timing, a different prompt source, and a smaller labeled pool. Notably, the LP-FT variant, which outperformed plain PRISM in LvL UP, falls below it on HAPPY (0.6646 vs. 0.6929).

We consider this evaluation as a preliminary generalization evaluation: these results suggest that the learned representations carry across cohorts: the same SSL-pretrained backbone, fine-tuned on HAPPY labels, continues to outperform the binary baselines. At the same time, the reversal between LP-FT and plain PRISM indicates that transfer strategies may be sensitive to dataset scale and cohort differences. We discuss this observation further in Section 7.3.

7 DISCUSSION

We discuss the implications of PRISM's distributional output along two axes. Section 7.1 covers the theoretical and methodological consequences of modeling receptivity as a time-to-event distribution, and Section 7.2 translates these into practical implications for JITAI system design.

7.1 Theoretical and Methodological Implications

The transition from binary classification to a time-to-event framework changes how we operationalize receptivity. By leveraging a Probability Mass Function (PMF) rather than a static decision, the architecture improves the transparency and debuggability of the model, allowing researchers to inspect the 'shape' of user urgency. As illustrated in Figure 3, this granular view reveals behavioral dynamics that are typically obscured by scalar metrics. For instance, we observed a distinct morphological shift in probability distributions between the early, mid and late phases of the intervention. This visual evidence quantitatively captures the phenomenon of 'prompt fatigue,' where the user's intent to respond persists, but the latency increases. Such insights transform receptivity from a black-box prediction into an interpretable behavioral metric, enabling a deeper understanding of how intervention fatigue manifests over longitudinal deployments.

The integration of SSL has implications for the robustness and data efficiency of mHealth modeling. Our results demonstrate that the "cold start" problem can be effectively mitigated by decoupling representation learning from task-specific discrimination. By pre-training on unlabeled sensor traces, the model acts as a stabilizer, learning robust representations of habitual behavior before encountering scarce receptivity labels. This suggests that future mHealth frameworks need not rely exclusively on expensive, high-burden labeled datasets. Instead, they can leverage the abundance of passive, unlabeled sensor data to build robust feature extractors that generalize better on unseen populations.

Finally, the architectural choices underlying this framework offer high reusability for the broader Ubiquitous Computing and HCI communities. The data footprint is not unique to receptivity; it is intrinsic to numerous predictive tasks. The PatchTST backbone keeps this capability computationally tractable: by attending over patches rather than raw timesteps, it reduces the memory cost of attention by a quadratic factor in the patch stride. Consequently, this architecture provides a transferable template for modeling other such context-dependent latency in mobile sensing. We emphasize that while the individual components (PatchTST, DeepHit, and self-supervised pretraining) are established, their integration in PRISM is driven by the structural constraints of receptivity modeling in mHealth. These include gradient instability arising from minute-level behavioral pairwise comparisons, heterogeneous sensor modalities with distinct statistical profiles, and label scarcity in longitudinal deployments. The design choices in PRISM follow directly from these constraints: the Safe ranking penalty stabilizes optimization under high-frequency comparisons, channel-isolated encoding accommodates heterogeneous sensor streams, and split-source self-supervised pretraining addresses the asymmetry between scarce labeled interactions and abundant passive sensing data. These domain-specific design adaptations make the time-to-event formulation viable on mobile sensing data. We anticipate these design principles could extend to other mobile sensing prediction tasks that share similar constraints.

7.2 Practical Implications for JITAI Systems

A distributional output provides a richer scheduling primitive than a binary flag. Instead of deciding only *whether* to trigger an intervention, a scheduler can use the predicted density to reason about *when* to deliver a prompt (Section 7.2.1), *what form* of intervention is appropriate (Section 7.2.2), and how these decisions shape the longer-term *relationship* with the user (Section 7.2.3). We develop each of these readings below.

7.2.1 From Reactive Polling to Proactive Scheduling. Current receptivity modules commonly operate on a reactive poll-and-trigger logic: the system queries a binary classifier at fixed intervals (e.g., every five minutes for up to

thirty minutes after a trigger event in Ally+ [29]) and fires the intervention on the first positive prediction. This design has been effective in establishing that receptivity can be predicted from mobile sensing. PRISM enables a complementary approach. By producing a full density over the actionable horizon in a single inference pass, the scheduler can identify the most probable response window directly and schedule the prompt accordingly, reserving re-inference for events that meaningfully change user context (e.g., phone unlock, location transition). Retry logic also becomes more principled: rather than re-polling, the scheduler can consult the cumulative distribution to determine whether a deferred slot still carries sufficient probability mass to justify delivery.

7.2.2 Matching Intervention Burden to Predicted Capacity. The shape of the density also informs *what kind* of intervention to deliver. Effective JITAIs often demand meaningful user effort: executing a target behavior can be "time-consuming, effortful, and even socially-deviant" [6], and some designs deliberately introduce interaction friction, such as Time2Stop's typing task that engages analytical reasoning [37]. A binary flag cannot distinguish a user who is available for the next minute from one who is available for the next hour; both register as receptive. PRISM's density can. When probability mass is concentrated near the prompt, the scheduler has evidence that a higher-burden synchronous intervention (e.g., a short EMA or guided task) is feasible. When mass is shifted to later bins, the same prediction supports a lower-burden asynchronous delivery (e.g., deferred notification or dashboard update) that can be consumed without interrupting ongoing activity.

7.2.3 User Experience and the Working Alliance. Over time, scheduling decisions accumulate into a relationship between the user and the system. A scheduler that only triggers when immediate responses are likely samples users in a narrow slice of their day, potentially biasing both the collected data and the user experience [17]. Binary systems that compensate with aggressive re-polling introduce a second cost: when support arrives at moments the user lacks capacity to engage, "well-intended support becomes more harmful than good" [15]. The soft-retry behavior enabled by PRISM addresses both issues. By treating delayed responses as predicted availability rather than rejection, the system can defer delivery when appropriate and avoid unnecessary re-polling. This also enables multi-objective scheduling strategies—for example, selecting delivery times that maximize predicted receptivity while avoiding recently used intervals. Interpreting delayed engagement as capacity rather than refusal aligns directly with the notion of a working alliance [15], as the system adapts to the user's availability rather than demanding immediate attention.

7.3 Limitations and Future Work

Although the PRISM architecture admits multiple competing events, our evaluation retains the binary, fixed-window operationalization to preserve parity with prior JITAI baselines (Section 6.2), with the multi-horizon sweep (Section 6.3) extending that parity across $\tau \in \{15, 30, 60, 90, 120\}$ minutes. This necessarily evaluates a fundamentally distributional output through a thresholded lens, and a unified contract that scores models directly under richer formulations (competing risks, calibrated latency, scheduling utility at deployment) remains an open direction for future work.

A central motivation for adopting a survival-based framework is its inherent capacity to handle *competing risks*. In real-world deployment, non-response is not a monolithic state; there is a significant semantic distinction between a user who actively dismisses a notification (indicating unavailability or burden) and one who simply misses it (indicating a lack of attention). Operationalizing this distinction in evaluation requires labels that separate the two outcomes, which our protocol does not provide; we therefore frame competing-risks scoring as the natural next step that the present architecture is already equipped to support.

Our primary evaluation also relies on a single, fixed-time prompt regime (the 09:00 daily anchor on LvL UP, Section 4), chosen to hold delivery context constant so that the contribution of the distributional shift to receptivity modeling could be isolated from variation introduced by the prompt schedule itself. The HAPPY

transfer evaluation (Section 6.6) offers preliminary evidence that PRISM's behavior is not tied to this anchor: the same backbone, fine-tuned on HAPPY labels, leads the binary lineup on chatbot-initiated prompts that reach participants at varied times across the day. We read this as feasibility evidence rather than a definitive test, given HAPPY's modest cohort size ($N = 865$ across 50 participants). A comprehensive evaluation across the broader range of delivery schedules used in deployed JITAIs remains an avenue for future work.

Additionally, we acknowledge that our discrete-time approach still imposes a temporal boundary via the 180-minute 'actionable horizon.' While we critique the rigidity of standard 10-minute binary windows, our model is similarly bounded by this architectural choice. The distinction, however, lies in the resolution: unlike binary classification, which collapses the timeline into a single label, our framework provides a fine-grained probability mass over discrete time bins *within* this horizon. This allows for flexible, post-hoc thresholding, though future work could explore continuous-time formulations to eliminate fixed horizons entirely.

Finally, the datasets that anchor both pre-training and evaluation share important constraints. Although our self-supervised pre-training drew on unlabeled data from distinct study cohorts, all collection occurred within Singapore and through a single application and sensing pipeline; the learned "behavioral grammars" may therefore encode local environmental biases (urban mobility patterns, tropical weather baselines) and pipeline-specific features that we cannot separate from the underlying behavior. The evaluation with the HAPPY study (Section 6.6) provides preliminary evidence of generalization under a different cohort and prompt regime, while holding the application, sensing pipeline, and feature schema fixed. PRISM maintains its relative advantage over the binary baselines at the 15-minute cutoff, suggesting that the learned representations carry across cohorts within this shared setting. At the same time, we observe that the LP-FT variant, which outperforms plain PRISM on LvL UP, does not transfer as consistently on HAPPY. We attribute this to a combination of smaller labeled data, distribution shift between source and target, and sensitivity of the transfer regime to source-scale calibration, though these factors cannot be fully disentangled with the available data. We therefore treat HAPPY as a transfer check on generalization rather than a basis for broad generalization claims. Future work can explore more robust evaluations for generalization by validating across diverse cultural, geographic, and application settings.

8 CONCLUSION

In this work, we propose PRISM, a framework that models receptivity as a probabilistic time-to-event distribution derived from longitudinal mobile sensing data. We validated its effectiveness on existing receptivity benchmarks, where it demonstrated competitive predictive fidelity and label efficiency in limited-label regimes. Beyond performance metrics, PRISM offers a robust and explainable pathway for modeling receptivity: through its self-supervised backbone, it effectively leverages unlabeled passive sensing data to address cold-start challenges, while its probabilistic output offers an expandable framework that better aligns with the continuous dynamics of user behavior. However, challenges remain, and future research should focus on validating transferability across diverse populations and exploring varied intervention delivery mechanisms. PRISM helps bridge the gap between static classification and granular behavioral modeling, paving the way for the next generation of proactive Just-in-Time Adaptive Interventions.

Acknowledgments

The research was conducted at the Future Health Technologies at the Singapore-ETH Centre, which was established collaboratively between ETH Zürich and the National Research Foundation Singapore. This research is supported by the National Research Foundation Singapore (NRF) under its Campus for Research Excellence and Technological Enterprise (CREATE) programme, and by the National Science Foundation (award no. IIS-2442593).

We thank Prof. Johan Gunnar Eriksson and the HAPPY study team (Yong Loo Lin School of Medicine, National University of Singapore) for providing access to the HAPPY study dataset.

The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the sponsors. Any mention of specific companies or products does not imply any endorsement by the authors, by their employers, or by the sponsors.

Conflicts of Interest

Samarth Negi, Tobias Kowatsch, Jacqueline L. Mair, and Florian von Wangenheim are affiliated with the Center for Digital Health Interventions (CDHI), a joint initiative of the Institute for Implementation Science in Health Care, University of Zurich; the Department of Management, Technology, and Economics, ETH Zurich; and the Institute of Technology Management and School of Medicine, University of St. Gallen. CDHI is funded in part by CSS, a Swiss health insurer; Mavie Next, an Austrian health insurer; and MTIP, a Swiss digital health investor. However, none of these companies was involved in this study. Tobias Kowatsch is cofounder of Pathmate Technologies, a university spinoff that offers digital health interventions. As of the end of 2023, Tobias Kowatsch no longer held any shares in or formal role with Pathmate Technologies. Pathmate Technologies was also not involved in this study.

References

[1] Hussein Ahmad Ahmad, Seyyed Kasra Mortazavi, Mohamed El Bahnasawi, Fadi Al Machot, Witesyavwirwa Vianney Kambale, and Kyandoghere Kyamakya. 2024. Enhanced Time Series Forecasting: Integrating PatchTST with BERT Layers. In *2024 International Conference on Applied Mathematics & Computer Science (ICAMCS)*. IEEE, Venice, Italy, 60–65. doi:10.1109/ICAMCS62774.2024.00014

[2] Daniel Avrahami and Scott E Hudson. 2006. Responsiveness in Instant Messaging: Predictive Models Supporting Inter-Personal Communication. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2006). doi:10.1145/1124772

[3] Jonathan Buckley and Ian James. 1979. Linear regression with censored data. *Biometrika* 66, 3 (1979), 429–436. doi:10.1093/biomet/66.3.429

[4] Oscar Castro, Jacqueline Louise Mair, Alicia Salamanca-Sanabria, Aishah Alattas, Roman Keller, Shenglin Zheng, Ahmad Jabir, Xiaowen Lin, Bea Franziska Frese, Chang Siang Lim, et al. 2023. Development of "LvL UP 1.0": a smartphone-based, conversational agent-delivered holistic lifestyle intervention for the prevention of non-communicable diseases and common mental disorders. *Frontiers in Digital Health* 5 (2023), 1039171.

[5] Oscar Castro, Jacqueline Louise Mair, Shenglin Zheng, Sarah Yi Xuan Tan, Ahmad Ishqi Jabir, Xiaoxi Yan, Bibhas Chakraborty, E Shyong Tai, Rob M. Van Dam, Florian Von Wangenheim, Elgar Fleisch, Konstadina Griva, Tobias Kowatsch, and Falk Müller-Riemenschneider. 2025. The LvL UP trial: Protocol for a sequential, multiple assignment, randomised controlled trial to assess the effectiveness of a blended mobile lifestyle intervention. *Contemporary Clinical Trials* 150 (March 2025), 107833. doi:10.1016/j.cct.2025.107833

[6] Woohyeok Choi, Sangkeun Park, Duyeon Kim, Youn kyung Lim, Uichin Lee, ChoiWoohyeok, ParkSangkeun, KimDuyeon, LimYoun-kyung, LeeUichin, Woohyeok Choi, and Uichin Lee. 2019. Multi-Stage Receptivity Model for Mobile Just-In-Time Health Intervention. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3 (6 2019), 1–26. Issue 2. doi:10.1145/3328910

[7] Shane P. Cross and Mario Alvarez-Jimenez. 2024. The digital cumulative complexity model: a framework for improving engagement in digital mental health interventions. *Frontiers in Psychiatry* 15 (Sept. 2024), 1382726. doi:10.3389/fpsyt.2024.1382726

[8] Ozlem Durmaz Incel and Sevda Özge Bursa. 2023. On-Device Deep Learning for Mobile and Wearable Sensing Applications: A Review. *IEEE Sensors Journal* 23, 6 (March 2023), 5501–5512. doi:10.1109/JSEN.2023.3240854

[9] Sannara EK, François Portet, and Philippe Lalanda. 2023. Transformer-based Models to Deal with Heterogeneous Environments in Human Activity Recognition. *Personal and Ubiquitous Computing* 27, 6 (Dec. 2023), 2267–2280. doi:10.1007/s00779-023-01776-3 arXiv:2209.11750 [cs].

[10] Joel E Fischer, Chris Greenhalgh, and Steve Benford. 2011. Investigating Episodes of Mobile Phone Activity as Indicators of Opportune Moments to Deliver Notifications. (2011).

[11] Yanlong Gao, Feng Xing, Lipeng Kang, Mingming Zhang, and Caiyan Qin. 2024. Ultra-Short-Term Wind Power Prediction Based on the ZS-DT-PatchTST Combined Model. *Energies* 17, 17 (Aug. 2024), 4332. doi:10.3390/en17174332

[12] Ting-Chen Chloe Hsu, Pauline Whelan, Julie Gandrup, Christopher J. Armitage, Lis Cordingley, and John McBeth. 2025. Personalized interventions for behaviour change: A scoping review of just-in-time adaptive interventions. *British Journal of Health Psychology* 30, 1 (Feb. 2025), e12766. doi:10.1111/bjhp.12766

[13] Xinyu Huang, Jun Tang, and Yongming Shen. 2024. Long time series of ocean wave prediction based on PatchTST model. *Ocean Engineering* 301 (June 2024), 117572. doi:10.1016/j.oceaneng.2024.117572

[14] I. R. James and P. J. Smith. 1984. Consistency Results for Linear Regression with Censored Data. *The Annals of Statistics* 12, 2 (June 1984). doi:10.1214/aos/1176346507

[15] Roman Keller, Florian V. Wangenheim, Jacqueline Mair, and Tobias Kowatsch. 2023. Receptivity to mobile health interventions. In *Digital Therapeutics for Mental Health and Addiction*. Elsevier, 65–77. doi:10.1016/B978-0-323-90045-4.00006-X

[16] Taesung Kim, Jinhee Kim, Yunwon Tae, Cheonbok Park, Jang-Ho Choi, and Jaegul Choo. 2022. REVERSIBLE INSTANCE NORMALIZATION FOR ACCURATE TIME-SERIES FORECASTING AGAINST DISTRIBUTION SHIFT. (2022).

[17] Zachary D King, Han Yu, Thomas Vaessen, Inez Myin-Germeys, and Akane Sano. 2024. Investigating Receptivity and Affect Using Machine Learning: Ecological Momentary Assessment and Wearable Sensing Study. *JMIR mHealth and uHealth* 12 (Feb. 2024), e46347. doi:10.2196/46347

[18] Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. 2022. Fine-Tuning can Distort Pretrained Features and Underperform Out-of-Distribution. doi:10.48550/arXiv.2202.10054 arXiv:2202.10054 [cs].

[19] Florian Künzler, Varun Mishra, Jan-Niklas Kramer, David Kotz, Elgar Fleisch, and Tobias Kowatsch. 2019. Exploring the state-of-receptivity for mhealth interventions. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 4 (2019), 1–27.

[20] Quang-Hung Le, Brijesh Patel, Donald Adjeroh, Gianfranco Doretto, and Ngan Le. 2025. SSL-SurvFormer: A Self-Supervised Learning and Continuously Monotonic Transformer Network for Missing Values in Survival Analysis. *Informatics* 12, 1 (March 2025), 32. doi:10.3390/informatics12010032

[21] Changhee Lee, Jinsung Yoon, and Mihaela Van Der Schaar. 2020. Dynamic-DeepHit: A Deep Learning Approach for Dynamic Survival Analysis With Competing Risks Based on Longitudinal Data. *IEEE Transactions on Biomedical Engineering* 67, 1 (Jan. 2020), 122–133. doi:10.1109/TBME.2019.2909027

[22] Changhee Lee, William Zame, Jinsung Yoon, and Mihaela Van Der Schaar. 2018. DeepHit: A Deep Learning Approach to Survival Analysis With Competing Risks. *Proceedings of the AAAI Conference on Artificial Intelligence* 32, 1 (April 2018). doi:10.1609/aaai.v32i1.11842

[23] Yingxue Li, Xiaolu Zhang, Jun Hu, Wenwen Xia, Ziyi Liu, Xiaobo Qin, Binjie Fei, and Jun Zhou. 2025. SurvFormer: A Transformer Based Framework for Survival Analysis in Insurance Underwriting. In *Companion Proceedings of the ACM on Web Conference 2025*. ACM, Sydney NSW Australia, 325–333. doi:10.1145/3701716.3715262

[24] Chao Liu, Ryen W. White, and Susan Dumais. 2010. Understanding web browsing behaviors through Weibull analysis of dwell time. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*. ACM, Geneva Switzerland, 379–386. doi:10.1145/1835449.1835513

[25] Jacqueline L. Mair, Ahmad I. Jabir, Alicia Salamanca-Sanabria, Oscar Castro, Shenglin Zheng, Roman Keller, Bea Franziska Frese, Chang Siang Lim, Aishah Alattas, Samarth Negi, Akshaye Shenoi, Rob M. Van Dam, EShyong Tai, Elgar Fleisch, Florian Von Wangenheim, Lorainne Tudor Car, Falk Müller-Riemenschneider, and Tobias Kowatsch. 2025. Feasibility of the LvL UP digital lifestyle coaching intervention designed to prevent non-communicable diseases and common mental disorders. *Scientific Reports* 16, 1 (Dec. 2025), 1243. doi:10.1038/s41598-025-30960-z

[26] Afra Mashhadi, Akhil Mathur, and Fahim Kawsar. 2014. The myth of subtle notifications. *UbiComp 2014 - Adjunct Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (2014), 111–114. doi:10.1145/2638728.2638759

[27] Abhinav Mehrotra, Mirco Musolesi, Robert Hendley, and Veljko Pejovic. 2015. Designing content-driven intelligent notification mechanisms for mobile applications. *UbiComp 2015 - Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (9 2015), 813–824. doi:10.1145/2750858.2807544

[28] Abhinav Mehrotra, Veljko Pejovic, Jo Vermeulen, Robert Hendley, and Mirco Musolesi. 2016. My phone and me: Understanding people's receptivity to mobile notifications. *Conference on Human Factors in Computing Systems - Proceedings* (5 2016), 1021–1032. doi:10.1145/2858036.2858566

[29] Varun Mishra, Florian Künzler, Jan-Niklas Kramer, Elgar Fleisch, Tobias Kowatsch, and David Kotz. 2021. Detecting Receptivity for mHealth Interventions in the Natural Environment. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 2 (June 2021), 1–24. doi:10.1145/3463492

[30] Varun Mishra, Florian Künzler, Jan-Niklas Kramer, Elgar Fleisch, Tobias Kowatsch, and David Kotz. 2021. Detecting Receptivity for mHealth Interventions in the Natural Environment. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 2 (2021), Article 74. doi:10.1145/3463492

[31] Leanne G. Morrison, Charlie Hargood, Veljko Pejovic, Adam W. A. Geraghty, Scott Lloyd, Natalie Goodman, Danius T. Michaelides, Anna Weston, Mirco Musolesi, Mark J. Weal, and Lucy Yardley. 2017. The Effect of Timing and Frequency of Push Notifications on Usage of a Smartphone-Based Stress Management Intervention: An Exploratory Trial. *PLOS ONE* 12, 1 (2017), e0169162. doi:10.1371/journal.pone.0169162

[32] I. Nahum-Shani, E. B. Hekler, and D. Spruijt-Metz. 2015. Building health behavior models to guide the development of just-in-time adaptive interventions: A pragmatic framework. *Health Psychol* 34s, 0 (2015), 1209–19. doi:10.1037/hea0000306

[33] Inbal Nahum-Shani, Shawna N. Smith, Bonnie J. Spring, Linda M. Collins, Katie Witkiewitz, Ambuj Tewari, and Susan A. Murphy. 2018. Just-in-time adaptive interventions (JITAIs) in mobile health: Key components and design principles for ongoing health behavior support. *Annals of Behavioral Medicine* 52 (5 2018), 446–462. Issue 6. doi:10.1007/s12160-016-9830-8

[34] Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. 2023. A Time Series is Worth 64 Words: Long-term Forecasting with Transformers. doi:10.48550/arXiv.2211.14730 arXiv:2211.14730 [cs].

[35] Mikio Obuchi, Jin Nakazawa, Wataru Sasaki, Hideyuki Tokuda, and Tadashi Okoshi. 2016. Investigating Interruptibility at Activity Breakpoints using Smartphone Activity Recognition API. *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct* (2016). doi:10.1145/2968219

[36] Tadashi Okoshi, Kota Tsubouchi, Masaya Taji, Takanori Ichikawa, and Hideyuki Tokuda. 2017. Attention and engagement-awareness in the wild: A large-scale study with adaptive notifications. *2017 IEEE International Conference on Pervasive Computing and Communications, PerCom 2017* (5 2017), 100–110. doi:10.1109/PERCOM.2017.7917856

[37] Adiba Orzikulova, Han Xiao, Zhipeng Li, Yukang Yan, Yuntao Wang, Yuanchun Shi, Marzyeh Ghassemi, Sung-Ju Lee, Anind K. Dey, and Xuhai "Orson" Xu. 2024. Time2Stop: Adaptive and Explainable Human-AI Loop for Smartphone Overuse Intervention. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–20. doi:10.1145/3613904.3642747 arXiv:2403.05584 [cs].

[38] Veljko Pejovic and Mirco Musolesi. 2014. InterruptMe: Designing intelligent prompting mechanisms for pervasive applications. *UbiComp 2014 - Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (2014), 897–908. doi:10.1145/2632048.2632062

[39] Olga Perski, Emily T. Hébert, Felix Naughton, Eric B. Hekler, Jamie Brown, and Michael S. Businelle. 2022. Technology-mediated just-in-time adaptive interventions (JITAIs) to reduce harmful substance use: a systematic review. *Addiction* 117, 5 (May 2022), 1220–1241. doi:10.1111/add.15687

[40] Martin Pielot, Bruno Cardoso, Kleomenis Katevas, Joan Serrà, Aleksandar Matic, and Nuria Oliver. 2017. Beyond Interruptibility: Predicting Opportune Moments to Engage Mobile Phone Users. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol* 1 (2017), 91. Issue 91. doi:10.1145/3130956

[41] Martin Pielot, Tilman Dingler, Jose San Pedro, and Nuria Oliver. 2015. When Attention is not Scarce - Detecting Boredom from Mobile Phone Usage. *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing - UbiComp '15* (2015). doi:10.1145/2750858

[42] Nilesh Kumar Sahu, Aditya Sneh, Snehil Gupta, and Haroon R. Lone. 2025. Real-World Receptivity to Adaptive Mental Health Interventions: Findings from an In-the-Wild Study. doi:10.48550/arXiv.2508.02817 arXiv:2508.02817 [cs].

[43] Alicia Salamanca-Sanabria, Seaw Jia Liew, Jacqueline Mair, Maria De Iorio, Young Doris Yee Ling, Mya Thway Tint, Yew Tong Wei, Karen Lim, Desmond Ong, Yu Chung Chooi, Vicky Tay, and Johan Gunnar Eriksson. 2024. A holistic lifestyle mobile health intervention for the prevention of type 2 diabetes and common mental disorders in Asian women with a history of gestational diabetes: a randomised control trial with 3-year follow-up protocol. *Trials* 25, 1 (July 2024), 443. doi:10.1186/s13063-024-08247-x

[44] Hillol Sarker, Moushumi Sharmin, Amin Ahsan Ali, Md Mahbubur Rahman, Rummana Bari, Syed Monowar Hossain, and Santosh Kumar. 2014. Assessing the availability of users to engage in just-in-time intervention in the natural environment. *UbiComp 2014 - Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (2014), 909–920. doi:10.1145/2632048.2636082

[45] Chen-Hsuan (Iris) Shih, Naofumi Tomita, Yanick X. Lukic, Álvaro Hernández Reguera, Elgar Fleisch, and Tobias Kowatsch. 2019. Breeze: Smartphone-based Acoustic Real-time Detection of Breathing Phases for a Gamified Biofeedback Breathing Training. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 4 (Dec. 2019), 1–30. doi:10.1145/3369835

[46] Gisbert W. Teepe, Ashish da Fonseca, Birgit Kleim, Nicholas C. Jacobson, Alicia Salamanca Sanabria, Lorainne Tudor Car, Elgar Fleisch, and Tobias Kowatsch. 2021. Just-in-time adaptive mechanisms of popular mobile apps for individuals with depression: Systematic app search and literature review. *Journal of Medical Internet Research* 23 (2021), e29412. Issue 9. doi:10.2196/29412

[47] Claire R. Van Genugten, Melissa S. Y. Thong, Wouter Van Ballegooijen, Annet M. Kleiboer, Donna Spruijt-Metz, Arnout C. Smit, Mirjam A. G. Sprangers, Yannik Terhorst, and Heleen Riper. 2025. Beyond the current state of just-in-time adaptive interventions in mental health: a qualitative systematic review. *Frontiers in Digital Health* 7 (Jan. 2025), 1460167. doi:10.3389/fdgth.2025.1460167

[48] Liyuan Wang and Lynn Carol Miller. 2020. Just-in-the-Moment Adaptive Interventions (JITAI): A Meta-Analytical Review. *Health Communication* 35, 12 (2020), 1531–1544. doi:10.1080/10410236.2019.1652388

[49] Liyuan Wang and Lynn Carol Miller. 2020. Just-in-the-Moment Adaptive Interventions (JITAI): A Meta-Analytical Review. *Health Communication* 35, 12 (Oct. 2020), 1531–1544. doi:10.1080/10410236.2019.1652388

[50] Zhiyuan Wang, Haoyi Xiong, Jie Zhang, Sijia Yang, Mehdi Boukhechba, Daqing Zhang, Laura E. Barnes, and Dejing Dou. 2022. From Personalized Medicine to Population Health: A Survey of mHealth Sensing Techniques. *IEEE Internet of Things Journal* 9, 17 (Sept. 2022), 15413–15434. doi:10.1109/JIOT.2022.3161046

[51] Guangyu Yang, Efrem Ghebreab, Jiaxi Xu, Xianen Qiu, Yiping Yuan, and Wensheng Sun. 2023. DeepPvisit: A Deep Survival Model for Notification Management. In *Companion Proceedings of the ACM Web Conference 2023*. ACM, Austin TX USA, 973–977. doi:10.1145/3543873.3587666

[52] Linyun Yu, Peng Cui, Chaoming Song, Tianyang Zhang, and Shiqiang Yang. 2017. A Temporally Heterogeneous Survival Framework with Application to Social Behavior Dynamics. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, Halifax NS Canada, 1295–1304. doi:10.1145/3097983.3098189

[53] Yiping Yuan, Jing Zhang, Shaunak Chatterjee, Shipeng Yu, and Romer Rosales. 2019. A State Transition Model for Mobile Notifications via Survival Analysis. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. ACM, Melbourne VIC Australia, 123–131. doi:10.1145/3289600.3290981

[54] Shenglin Zheng, Oscar Castro, Jacqueline Louise Mair, Ahmad I. Jabir, Sarah Yi Xuan Tan, Akshaye Shenoi, Samarth Negi, Ruth Rachael Mathews, Rachel Woon Sim Koh, Xiaoxi Yan, Bibhas Chakraborty, E. Shyong Tai, Rob M. Van Dam, Florian Von Wangenheim, Elgar Fleisch, Konstadina Griva, Tobias Kowatsch, and Falk Müller-Riemenschneider. 2026. Feasibility of the "LvL UP" trial: a pilot sequential multiple assignment randomised trial of an adaptive, holistic mHealth lifestyle coaching intervention. *International Journal of Behavioral Nutrition and Physical Activity* 23, 1 (Jan. 2026), 16. doi:10.1186/s12966-025-01869-7