Constructing the Mental Health Phenome: An Open Multimodal Dataset Linking Digital Behavior, Physical Health, and Mental Wellbeing

Shakson Isaac

Department of Biomedical Informatics
Harvard Medical School
shakson_isaac@g.harvard.edu

Ambika Grover

Department of Biomedical Informatics
Harvard Medical School
ambikagrover@college.harvard.edu

Yentl Collin
Department of Biomedical Informatics
Harvard Medical School
yentl.collin@eleves.enpc.fr

John Torous
Division of Digital Psychiatry
Beth Israel Deaconess Medical Center
jtorous@bidmc.harvard.edu

Chirag J. Patel
Department of Biomedical Informatics
Harvard Medical School
chirag_patel@hms.harvard.edu

Abstract

Mental health, despite its global importance, lacks a large-scale, multimodal, open dataset. We propose the **Mental Health Phenome**: an openly shareable collection of longitudinal smartphone behavior, wearable physiology, social media activity, and mental health measures. Enabling foundation models for predictive, interpretable, and personalized discovery across psychiatry, public health, and digital well-being. Starting with *deeply phenotyped* cohorts, the Mental Health Phenome is designed to scale to broader populations, combining mechanistic insights with global applicability.

Accepted to the 39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: AI for Science.

1 AI Task Definition

Mental health remains uniquely challenging to quantify compared to physical health. Unlike physical health, routinely measured via observable phenotypes like weight or blood pressure, mental health relies on sparse, episodic clinical data. In the absence of frequent longitudinal touch points, there is much room for discovery in the mental health space: spotting early signs of depressive relapse, tracking mood fluctuations via passive data streams (sleep, mobility patterns), or detecting new behavioral markers of well-being from digital interactions and daily routines. Therefore, accurate diagnosis and effective long-term management, particularly for complex or severe conditions, requires individualized, person-centered care, enabling study of the effects of digital behavior on mental health.

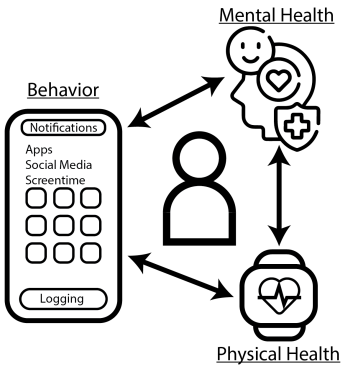

Figure 1: Mental Health Phenome: Behaviors captured via smartphones influence mental and physical health.

Our *Mental Health Phenome* (MHP) dataset (Figure 1) will address this gap by enabling and promoting the following AI tasks. These tasks are designed for the individual participants, with clinicians and researchers as secondary users. **Prediction:** Forecast mental health risks and disease trajectory (e.g., anxiety, depression, manic episodes) from longitudinal digital behaviors. **Representation learning:** Build foundation models linking digital behavior to mental health. **Interpretability:** Provide human-centered explanations of why specific digital behaviors correlate with both mental and physical health outcomes. Formally, the dataset will enable learning functions of the form:

$$\text{Mental State} \sim f(\text{digital behaviors, physiology, social context}).$$

2 Dataset Rationale

Prior Work and Feasibility: Existing resources are fragmented (sub-1,000 users, short follow-up, narrow demographic coverage) and fail to integrate multimodal behavioral, physiological, and clinical data at scale. GLOBEM [1] pioneered multi-year mobile and wearable sensing but is confined to short 10-week undergraduate cohorts with shallow clinical measures. ABCD [2] provides large-scale information in adolescents, but offers shallow wearable data and broad symptom checklists. AMP SCZ [3] is the closest precedent, integrating clinical, imaging, and digital health in schizophrenia and psychosis risk cohorts, but remains constrained by narrow scope, intensive protocols, and limited scalability. Collectively, these efforts demonstrate the promise of large-scale digital health resources but leave critical gaps: (i) limited multimodal integration, (ii) narrow populations, (iii) absence of cross-condition, clinically rich data, and (iv) use-cases that are primarily research-driven with few approaches that are amenable to "personalization".

To address this, the MHP begins with a *deeply phenotyped cohort* (e.g. bipolar disorder) and scales via widely deployed APIs, overcoming prior challenges of retention and scalability. For example, pilot studies on bipolar disorder (BD) suggest that it is often associated with changes in communication and irregular online behaviors [4]. From a deep phenotyping perspective, linking these behavioral patterns to physiological markers, such as heart rate variability, offers a promising avenue for uncovering the underpinnings of BD (see Appendix A.1).

Data types: As illustrated in Figure 2, MHP integrates active inputs (surveys, journaling) with passive streams (wearables, app usage, social media), plus demographic and clinical metadata (e.g. age, sex, ethnicity, geographic region, self-reported mental health, clinical indicators and assessments). Active features such as questionnaires and surveys will enable researchers to collect information about mood, thought patterns, medication history, and demographic information. Passive features from APIs such as phone logs (notifications, app use, screentime), social media engagement, and activity wearable signals (e.g. heart rate, sleep, steps) provide high-frequency multimodal time series.

Scale and resolution: To infer behavior-health links, thousands of users with high-frequency longitudinal measures are needed. Resolution spans coarse aggregates (daily screentime) to fine-grained behaviors and engagement (content type, time-of-day, scrolling behavior). Where

possible, we aim to achieve a high degree of specificity in our dataset either through APIs or direct partnerships (Section 4) as opposed to more generic data on screen time use.

3 Acceleration Potential

Recent work shows foundation models trained on behavioral wearable data can improve health prediction [5], but current datasets lack multimodal scale and clinical depth required with mental health conditions. MHP would unlock a new class of foundation models for mental health: multimodal embeddings that link digital behavior (e.g., phone use, activity, sleep, social media) with clinical outcomes. Such models could power early detection of relapse, personalized interventions, and digital mental health coaching. Beyond psychiatry, applications extend to public health, workplace well-being, and education. These data would become a commons for accelerating mental health research and public benefit (see Appendix A.2).

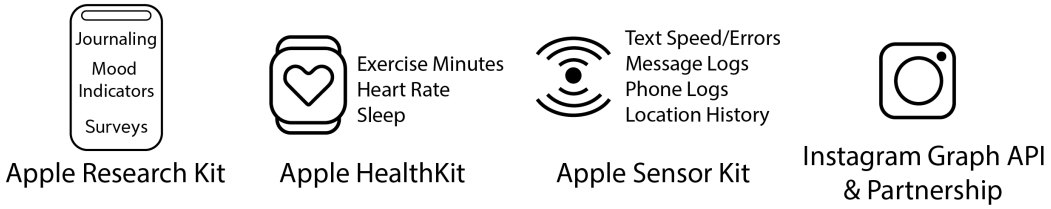

Figure 2: Proposed data pipeline integrating Apple APIs (HealthKit, SensorKit, ResearchKit) and Instagram Graph API, with informed consent and de-identification safeguards.

4 Data Creation Pathway

We leverage existing, widely deployed APIs such as Apple HealthKit, SensorKit, and ResearchKit [6][7][8] to provide continuous signals (heart rate, sleep, mobility, journaling, app usage). Instagram Graph API enables collection of consented digital social data [9], while the COS–Meta partnership demonstrates precedent for deeper, privacy-preserving access to Instagram behavioral data at research scale [10].

Standardized surveys will complement these digital measures with validated mental health instruments. All data will be collected under informed consent and IRB approval, with strict de-identification and secure storage. For sensitive modalities, federated learning pipelines [11] [12] will be explored to enable large-scale integration without centralizing raw data. For benchmarking and reproducibility, de-identified datasets will be available under controlled centralized access, while federated learning provides an additional privacy-preserving pathway for scaling to larger, real-world deployments. Further details on API usage and federated learning are provided in Appendix A.3 and Appendix A.4.

5 Cost and Scalability

The dataset is cost-feasible at scale (see Appendix A.5). A pilot cohort of 1,000–5,000 users can be launched for under $250K (server/API costs plus recruitment). Scaling to tens of thousands is achievable through academic consortia (e.g., ABCD, All of Us) and partnerships with platform providers. Because the data pipeline leverages existing devices and APIs, marginal costs per participant are low. Cloud-based, privacy-preserving infrastructure ensures that the dataset can grow sustainably without prohibitive overhead.

6 Vision

The Mental Health Phenome (MHP) is uniquely designed to balance depth and breadth: starting with condition-specific cohorts, scaling via widely deployed APIs, and embedding privacy and governance through design features, ensuring ethical scalability beyond psychiatry into public health, workplace well-being, and education. MHP will also establish foundation-model-ready benchmarks with standardized tasks, splits, and evaluation metrics. This combination of deep clinical grounding, scalable multimodal design, benchmark creation, and ethical governance sets MHP apart as the first truly foundational dataset for AI in mental health.

References

[1] Xuhai Xu, Han Zhang, Yasaman Sefidgar, Yiyi Ren, Xin Liu, Woosuk Seo, Jennifer Brown, Kevin Kuehn, Mike Merrill, Paula Nurius, Shwetak Patel, Tim Althoff, Margaret Morris, Eve Riskin, Jennifer Mankoff, and Anind Dey. Globem dataset: Multi-year datasets for longitudinal human behavior modeling generalization. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 24655–24692. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/9c7e8a0821dfcb58a9a83cbd37cc8131-Paper-Datasets_and_Benchmarks.pdf.

[2] Jason J Liu, Beatrice Borsari, Yunyang Li, Susanna X Liu, Yuan Gao, Xin Xin, Shaoke Lou, Matthew Jensen, Diego Garrido-Martín, Terril L Verplaetse, Garrett Ash, Jing Zhang, Matthew J Girgenti, Walter Roberts, and Mark Gerstein. Digital phenotyping from wearables using AI characterizes psychiatric disorders and identifies genetic associations. *Cell*, 188(2):515–529.e15, January 2025.

[3] Johanna T W Wigman, Ann Ee Ching, Yoonho Chung, Habiballah Rahimi Eichi, Erlend Lane, Carsten Langholm, Aditya Vaidyam, Andrew Jin Soo Byun, Anastasia Haidar, Jessica Hartmann, Angela Nunez, Dominic Dwyer, Adibah Amani Nasarudin, Owen Borders, Isabelle Scott, Zailyn Tamayo, Priya Matneja, Kang-Ik Cho, Jean Addington, Luis K Alameda, Celso Arango, Nicholas J K Breitborde, Matthew R Broome, Kristin S Cadenhead, Monica E Calkins, Eric Yu Hai Chen, Jimmy Choi, Philippe Conus, Cheryl M Corcoran, Barbara A Cornblatt, Covadonga M Diaz-Caneja, Lauren M Ellman, Paolo Fusar-Poli, Pablo A Gaspar, Carla Gerber, Louise Birkedal Glenthøj, Leslie E Horton, Christy Lai Ming Hui, Joseph Kambeitz, Lana Kambeitz-Ilankovic, Matcheri S Keshavan, Sung-Wan Kim, Nikolaos Koutsouleris, Kerstin Langbein, Daniel Mamah, Daniel H Mathalon, Vijay A Mittal, Merete Nordentoft, Godfrey D Pearlson, Jesus Perez, Diana O Perkins, Albert R Powers, 3rd, Jack Rogers, Fred W Sabb, Jason Schiffman, Jai L Shah, Steven M Silverstein, Stefan Smesny, Walid Yassin, William S Stone, Gregory P Strauss, Judy L Thompson, Rachel Upthegrove, Swapna Verma, Jijun Wang, Daniel H Wolf, Phillip Wolff, Accelerating Medicines Partnership® Schizophrenia (AMP® SCZ), Laura M Rowland, Simon D'Alfonso, Ofer Pasternak, Sylvain Bouix, Patrick D McGorry, Rene S Kahn, John M Kane, Carrie E Bearden, Scott W Woods, Martha E Shenton, Barnaby Nelson, Justin T Baker, and John Torous. Digital health technologies in the accelerating medicines partnership® schizophrenia program. *Schizophrenia (Heidelb.)*, 11(1):83, June 2025.

[4] Maria Faurholt-Jepsen, Maj Vinberg, Mads Frost, Sune Debel, Ellen Margrethe Christensen, Jakob E Bardram, and Lars Vedel Kessing. Behavioral activities collected through smartphones and the association with illness activity in bipolar disorder. *Int. J. Methods Psychiatr. Res.*, 25(4):309–323, December 2016.

[5] Eray Erturk, Fahad Kamran, Salar Abbaspourazad, Sean Jewell, Harsh Sharma, Yujie Li, Sinead Williamson, Nicholas J Foti, and Joseph Futoma. Beyond Sensor Data: Foundation Models of Behavioral Data from Wearables Improve Health Predictions, 2025. URL https://arxiv.org/abs/2507.00191.

[6] Apple Inc. HealthKit. https://developer.apple.com/healthkit/, . Accessed: 2025-08-24.

[7] Apple Inc. ResearchKit. https://researchkit.org/, . Accessed: 2025-08-24.

[8] Apple Inc. SensorKit. https://developer.apple.com/documentation/sensorkit, . Accessed: 2025-08-24.

[9] Meta Platforms, Inc. Instagram Graph API. https://developers.facebook.com/docs/instagram-api. Accessed: 2025-08-24.

[10] Center for Open Science. Instagram Data Access Pilot for Well-being Research. https://www.cos.io/meta. Accessed: 2025-08-24.

[11] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017. URL https://proceedings.mlr.press/v54/mcmahan17a/mcmahan17a.pdf.

[12] Peter Kairouz, Brendan Avent McMahan, et al. Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, 2021. URL https://arxiv.org/abs/1912.04977. Preprint version: arXiv:1912.04977.

[13] AI-READI Project (Bridge2AI, NIH). Artificial Intelligence-Ready and Exploratory Atlas for Diabetes Insights (AI-READI). https://aireadi.org/. Accessed: 2025-08-28.

[14] Lee Reicher, Smadar Shilo, Anastasia Godneva, Guy Lutsker, Liron Zahavi, Saar Shoer, David Krongauz, Michal Rein, Sarah Kohn, Tomer Segev, Yishay Schlesinger, Daniel Barak, Zachary Levine, Ayya Keshet, Rotem Shaulitch, Maya Lotan-Pompan, Matan Elkan, Yeela Talmor-Barkan, Yaron Aviv, Maya Dadiani, Yonatan Tsodyks, Einav Nili Gal-Yam, Haim Leibovitzh, Lael Werner, Roie Tzadok, Nitsan Maharshak, Shin Koga, Yulia Glick-Gorman, Chani Stossel, Maria Raitses-Gurevich, Talia Golan, Raja Dhir, Yotam Reisner, Adina Weinberger, Hagai Rossman, Le Song, Eric P. Xing, and Eran Segal. Deep phenotyping of health–disease continuum in the human phenotype project. *Nature Medicine*, 31(9):3191–3203, July 2025. ISSN 1546-170X. doi: 10.1038/s41591-025-03790-9. URL http://dx.doi.org/10.1038/s41591-025-03790-9.

[15] Tri Dao and Albert Gu. Transformers are SSMs: Generalized Models and Efficient Algorithms Through Structured State Space Duality, 2024. URL https://arxiv.org/abs/2405.21060.

[16] Yuwei Zhang, Kumar Ayush, Siyuan Qiao, A. Ali Heydari, Girish Narayanswamy, Maxwell A. Xu, Ahmed A. Metwally, Shawn Xu, Jake Garrison, Xuhai Xu, Tim Althoff, Yun Liu, Pushmeet Kohli, Jiening Zhan, Mark Malhotra, Shwetak Patel, Cecilia Mascolo, Xin Liu, Daniel McDuff, and Yuzhe Yang. SensorLM: Learning the Language of Wearable Sensors, 2025. URL https://arxiv.org/abs/2506.09108.

[17] Apple Inc. HealthKit Framework. https://developer.apple.com/documentation/healthkit, . Accessed: 2025-08-28.

[18] Google Inc. Health Connect API. https://developer.android.com/guide/health-and-fitness/health-connect, . Accessed: 2025-08-28.

[19] Sage Bionetworks and Cornell Tech. ResearchStack: An Open-Source Framework for Mobile Research on Android. http://researchstack.org/. Accessed: 2025-08-28.

[20] Google Inc. SensorManager | Android Developers. https://developer.android.com/reference/android/hardware/SensorManager, . Accessed: 2025-08-28.

[21] Thomas J Littlejohns, Jo Holliday, Lorna M Gibson, Steve Garratt, Niels Oesingmann, Fidel Alfaro-Almagro, Jimmy D Bell, Chris Boultwood, Rory Collins, Megan C Conroy, Nicola Crabtree, Nicola Doherty, Alejandro F Frangi, Nicholas C Harvey, Paul Leeson, Karla L Miller, Stefan Neubauer, Steffen E Petersen, Jonathan Sellors, Simon Sheard, Stephen M Smith, Cathie L M Sudlow, Paul M Matthews, and Naomi E Allen. The UK biobank imaging enhancement of 100,000 participants: rationale, data collection, management and future directions. *Nat. Commun.*, 11(1):2624, May 2020.

[22] Observational Health Data Sciences and Informatics (OHDSI). Data Standardization — OMOP Common Data Model. https://www.ohdsi.org/data-standardization/. Accessed: 2025-08-28.

[23] HL7 International. FHIR Overview. https://www.hl7.org/fhir/overview.html. Accessed: 2025-08-28.

[24] Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H. Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. Practical secure aggregation for privacy-preserving machine learning. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, CCS '17, page 1175–1191, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450349468. doi: 10.1145/3133956.3133982. URL https://doi.org/10.1145/3133956.3133982.

[25] Mang Ye, Wei Shen, Bo Du, Eduard Snezhko, Vassili Kovalev, and Pong C. Yuen. Vertical federated learning for effectiveness, security, applicability: A survey. *ACM Comput. Surv.*, 57(9), April 2025. ISSN 0360-0300. doi: 10.1145/3720539. URL https://doi.org/10.1145/3720539.

[26] Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, CCS'16, page 308–318. ACM, October 2016. doi: 10.1145/2976749.2978318. URL http://dx.doi.org/10.1145/2976749.2978318.

[27] Nicolas Papernot, Shuang Song, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Úlfar Erlingsson. Scalable Private Learning with PATE, 2018. URL https://arxiv.org/abs/1802.08908.

[28] Praneeth Vepakomma, Otkrist Gupta, Tristan Swedish, and Ramesh Raskar. Split learning for health: Distributed deep learning without sharing raw patient data, 2018. URL https://arxiv.org/abs/1812.00564.

[29] Hongxia Li, Zhongyi Cai, Jingya Wang, Jiangnan Tang, Weiping Ding, Chin-Teng Lin, and Ye Shi. Fedtp: Federated learning by transformer personalization. *IEEE Transactions on Neural Networks and Learning Systems*, 35(10):13426–13440, 2024. doi: 10.1109/TNNLS.2023.3269062.

[30] National Institutes of Health. All of us research program. Program overview, 2020. URL `https://allofus.nih.gov`.

[31] John Torous, Patrick Staples, Ian Barnett, Luis Sandoval, Matcheri Keshavan, and Jukka-Pekka Onnela. New tools for new research in psychiatry: A scalable and customizable platform to empower data driven smartphone research. *JMIR Mental Health*, 3(2):e16, 2016. doi: 10.2196/mental.5165. URL `https://mental.jmir.org/2016/2/e16/`.

[32] Deloitte Insights. Digital health: Navigating the challenges of always-on care. Industry report, 2021. URL `https://www2.deloitte.com/us/en/insights/industry/health-care/digital-health-always-on-care.html`.

A Technical Appendices and Supplementary Material

A.1 Deep Phenotype Model

As our proposed MHP intends to collect a vast quantity of data, we propose a deeply phenotyped cohort to rigorously assess how digital behaviors, physiology, and social context map to a "known" mental state. As an example of this in the physical health space, the AI-READI (Artificial Intelligence-Ready and Exploratory Atlas for Diabetes Insights) project consists of a dataset of 4,000 individuals for type 2 diabetes research. Participants completed both on-site assessments and at-home monitoring for approximately 10 days, where at-home data consisted of continuous glucose monitoring (CGM), physical activity monitoring, and in-home environmental sensor data [13]. The resulting dataset is ripe to train/test AI models on diabetes-related outcomes and empower precision care.

Using datasets such as AI-READI as inspiration, our deep phenotyped approach will select for one mental health-related disorder (e.g., bipolar disorder), and aim to provide a complete dataset ripe for discovery. This can both enable discovery as in the AI-READI dataset, and also provide valuable information on artifacts, noise, and signals that are relatively *unimportant* to map to mental health. Creating several of these cohorts can individually propel the field and enable valuable discovery.

Eventually, we hope to undertake a complete MHP that is analogous to the Human Phenotype Project (HPP), a deeply phenotyped longitudinal cohort with 28,000 enrolled participants. For this dataset, digital data streams included continuous glucose monitoring (CGM), sleep tracking, lifestyle, and nutrition logging over several weeks to untangle lifestyle data and general physical health outcomes. After collecting these data, the researchers designed foundation models for prediction and disease forecasting. We anticipate undertaking a similar yet more expansive approach for many mental health disorders as the subsequent step to our deep phenotyped datasets [14].

A.2 Model Acceleration

The acceleration potential of the MHP is marked by recent advances in foundation models for wearable sensor data. Whereas early work emphasized modeling raw sensor streams at second resolution, recent studies show that thoughtfully summarized physiological metrics, aggregated over days or weeks, can yield more accurate and interpretable predictions of disease outcomes and medication usage [5]. Utilizing the Mamba2 architecture [15], these derived metrics not only achieved strong performance but also reduced training complexity by lowering dimensionality and noise, underscoring that the design of informative behavioral summaries from raw signals is critical. In the MHP, such metrics will be systematically engineered to accelerate model training, reduce data requirements, and improve generalization across diverse clinical endpoints.

Complementarily, sensor-language alignment links time-series data with natural language, producing interpretable embeddings that connect digital behaviors to human-readable concepts [16]. By aligning time-series sensor data with language representations, such models facilitate the creation of embeddings that bridge digital behaviors and clinical concepts. In MHP, where sensor, smartphone, and validated clinical labels are co-located longitudinally, such alignment could enable zero-/few-shot symptom tagging, text-sensor retrieval, and clinician-facing summaries accelerating the translational potential of foundation models trained on MHP.

A.3 Details and Limitations of Proposed APIs

An important aspect of the proposed model suggests the use of existing APIs, such as Apple HealthKit, SensorKit, and ResearchKit. Herein, we will discuss the specific features that each API provides, as well as potential limitations associated with their use.

Apple's HealthKit API organizes data through an HKObjectType hierarchy, which provides a standardized way to integrate multi-modal health data across different applications and devices. Each health record is represented as a type of "sample" with subclasses that define how the data is stored and interpreted.

- Quantity Samples **(HKQuantitySample)** capture numeric values paired with units. These are used for continuous or countable measures such as height, weight, body mass index (BMI), step count, distance walked, calories burned, and heart rate.
- Category Samples **(HKCategorySample)** represent discrete states or events. For example, they can record phases of sleep (e.g., in-bed, asleep, awake), menstrual flow categories, or the presence of a symptom like a headache.
- Workout Objects **(HKWorkout)** bundle together data from exercise sessions, such as start and end times, activity type, total energy burned, distance, and associated heart rate samples.
- Correlation Objects **(HKCorrelation)** allow grouping of related samples into a single event, such as combining blood pressure systolic and diastolic measurements or food intake details.
- Clinical Records **(HKClinicalRecord)** enable integration of structured health records (e.g., lab results, immunizations, medications) from electronic health record (EHR) systems.
- Electrocardiograms **(HKElectrocardiogram)** and other specialized types capture data generated by Apple Watch sensors, such as ECG waveforms, heart rate variability, and VO2 max estimates [17].

Every sample in HealthKit also includes metadata such as the start and end date, source (app or device), and optional contextual information. In practice, this structure enables apps to both read and write to HealthKit, and for our MHP will allow for ease of use for interpretation by researchers. Of note, Google's Health Connect (the Android comparative) works using a very similar model, and can be incorporated or substituted as desired for wider use [18].

Meanwhile, **SensorKit** is focused on behavioral and environmental data streams. Some examples of data classes include:

- SRAmbientLightSample – brightness level.
- SRDeviceUsageReport – app usage time, notifications.
- SRAudioInputDeviceSample – microphone input info (not raw audio, but characteristics like power levels)
- SRMessagesUsageReport – messaging events (counts, not content).
- SRPhoneUsageReport – calls made, duration, etc.
- SRTouchEvent – taps, swipe gestures, typing cadence.

As before, each record includes the time window, sensor-specific fields (e.g. app name, battery), and metadata [8].

Finally, **Apple Research Kit** is an open-sourced framework deployed by researchers to build iOS apps for clinical studies. Embedded with it are a number of helpful features including:

- Built-in survey modules let you ask about mood, stress, sleep quality, symptoms at regular intervals.
- ResearchKit includes tasks like reaction time, tapping, spatial memory, Stroop tests.
- Ability to synthesize self-reports (mood surveys) with passive data (sleep, activity, heart rate from HealthKit; phone usage/typing cadence from SensorKit) [7].

Android's ResearchStack [19] is analogous to Apple Research Kit, and Android Sensor Manager is comparable to SensorKit [20].

When building an integrated dataset across HealthKit, SensorKit, and ResearchKit, limitations stem from differences in structure, timing, access, and completeness. Each framework produces different data types: continuous physiological time series from HealthKit, aggregated behavioral metadata from SensorKit, and discrete survey or task results from ResearchKit. Synchronizing data is also challenging since sampling frequencies and time may not be in line.

However, a useful solution to explore is mapping data into a shared common model in a way that is similar to the way in which the UK Biobank (UKB) [21] integrated multi-modal data. In the UKB, raw data come in many different formats: ICD codes (EHR), DICOMs imaging, Fitbit-like accelerometry, self-report survey fields. These are then harmonized into standardized fields (coding files, field IDs, controlled vocabularies). This is analogous to mapping HealthKit/SensorKit/ResearchKit into a common model like OMOP [22] or FHIR [23].

A.4 Data Privacy and Model Architecture

We will employ federated learning (FL) to ensure privacy-preserving AI development in the *Mental Health Phenome* dataset. FL keeps sensitive health and behavioral data on participants' devices while transmitting only model updates to a central server [11]. This approach minimizes privacy risks while enabling training across multiple individuals.

To strengthen privacy in FL, we will explore the following methods:

- **Secure Aggregation:** Ensures the server only observes the aggregate of client updates, preventing inference about any single participant's data [24].
- **Vertical Federated Learning (VFL):** Enables collaboration across institutions holding different features for the same individuals without sharing raw data [25].
- **Differential Privacy (DP):** Adds calibrated noise to training (e.g., DP-SGD [26], PATE [27]) to provide formal guarantees against membership and attribute inference.
- **Split Learning (SplitNN):** Clients transmit intermediate activations instead of raw data or gradients, with labels retained locally to enhance privacy [28].
- **Personalized Federated Learning:** Recent work such as FedTP trains personalized self-attention layers in Transformers under FL objectives, improving performance in heterogeneous (non-IID) data settings [29].

A.5 Cost and Scalability

The main costs of a pilot cohort derive from participant recruitment and incentives, together with cloud storage and secure data infrastructure. By leveraging participants' own smartphones and wearables through widely deployed APIs (e.g., HealthKit, ResearchKit, Instagram Graph API), the Mental Health Phenome minimizes device costs and achieves low marginal cost per participant. Based on precedents from digital phenotyping and population health cohorts, a pilot of 1,000–5,000 participants can be launched for under $250K, with scalability to tens of thousands enabled by academic consortia and platform partnerships (Table 1).

Table 1: **Estimated pilot costs for 1,000–5,000 participants.**

Category	**Description**	**Unit Cost (per user)**	**Total (1,000 users)**	**Total (5,000 users)**	**Relevant References**
Recruitment & Incentives	Participant Compensation and Outreach	$40–100	$ 40–100K	$ 200–500K	GLOBEM, BEIWE, DELOITTE
Participant Support	Communications and Tech Support	$10–20	$ 10–20K	$ 50–100K	BEIWE, DELOITTE
Server & Cloud Infrastructure	Data ingestion, storage, backups, secure cloud	$5–15	$ 5–15K	$ 25–75K	GLOBEM, DELOITTE
API / App Development	Integration of Apple APIs, Instagram API, survey platform	Fixed $ 30–50K	$ 30–50K	$ 30–50K	GLOBEM, BEIWE
Data Privacy & Security	Encryption, IRB-required de-identification, monitoring	Fixed $ 10–20K	$ 10–20K	$ 10–20K	AMP–SCZ, ALLOFUS
Regulatory & IRB	Protocol prep, renewals, legal review	Fixed $ 10K	$ 10K	$ 10K	AMP–SCZ, ALLOFUS
Total Estimate	—	—	$ **105–215K**	$ **325–755K**	—

Acronyms. ALLOFUS = NIH All of Us Research Program [30]; AMP–SCZ = Wigman et al. 2025 [3]; BEIWE = Torous et al. 2016 [31]; DELOITTE = Deloitte Digital Health Survey [32]; GLOBEM = Xu et al. 2022 [1].