Predictive Modeling of Mental Illness Onset Using Wearable Devices and Medical Examination Data: Machine Learning Approach

Tomoki Saito 1*, Hikaru Suzuki 1 and Akifumi Kishi 2*

1 JMDC Inc., Tokyo, Japan, 2 Graduate School of Education, The University of Tokyo, Tokyo, Japan

The prevention and treatment of mental illness is a serious social issue. Prediction and intervention, however, have been difficult because of lack of objective biomarkers for mental illness. The objective of this study was to use biometric data acquired from wearable devices as well as medical examination data to build a predictive model that can contribute to the prevention of the onset of mental illness. This was an observational study of 4,612 subjects from the health database of society-managed health insurance in Japan provided by JMDC Inc. The inputs to the predictive model were 3-months of continuous wearable data and medical examinations within and near that period; the output was the presence or absence of mental illness over the following month, as defined by insurance claims data. The features relating to the wearable data were sleep, activity, and resting heart rate, measured by a consumer-grade wearable device (specifically, Fitbit). The predictive model was built using the XGBoost algorithm and presented an area-under-the-receiver-operating-characteristic curve of 0.712 (SD = 0.02, a repeated stratified group 10-fold cross validation). The top-ranking feature importance measure was wearable data, and its importance was higher than the blood-test values from medical examinations. Detailed verification of the model showed that predictions were made based on disrupted sleep rhythms, mild physical activity duration, alcohol use, and medical examination data on disrupted eating habits as risk factors. In summary, the predictive model showed useful accuracy for grouping the risk of mental illness onset, suggesting the potential of predictive detection, and preventive intervention using wearable devices. Sleep abnormalities in particular were detected as wearable data 3 months prior to mental illness onset, and the possibility of early intervention targeting the stabilization of sleep as an effective measure for mental illness onset was shown.

Keywords: machine learning, medical examination, mental illness, mHealth, physical activity, predictive detection, sleep, wearable data

INTRODUCTION

Mental illness, including depression, is highly prevalent worldwide, and the lifetime prevalence has been reported to be as high as ∼50 and 20% in the United States and Japan, respectively (1, 2). Because mental illness causes a significant decrease in quality of life and socioeconomic loss, its treatment and prevention are among the most serious societal challenges of today (3). Prevention
and detection at the onset of mental illness are extremely important because of the generally low remission rate for mental illness and the favorable prognosis of early initiation of treatment (4, 5). Predictive detection is difficult, however, because there are no objective biomarkers for mental illness.

Meanwhile, the spread of wearable sensor technologies has enabled the continuous acquisition of detailed daily lifestyle data (e.g., sleep and activity) from individuals (6). Furthermore, advances in computational processing capabilities and machine learning have enabled the development of predictive models that utilize high-frequency and high-dimensional data linked from wearable devices. Therefore, research combining big health data collection, machine learning, and artificial intelligence (AI) analysis in the health care field has attracted attention, and there are high expectations for their application to the objective examination and onset prediction of mental illness (7–12).

Previous studies have primarily examined the usefulness of sleep and physical activity data. Fang et al. reported that irregular sleep rhythms were associated with worsening of mood and depressive symptoms the following day, indicating the importance of maintaining regular sleep habits for the maintenance and improvement of mental health (13). Nakamura et al. reported that time-series analysis of physical activity data could distinguish depression and schizophrenia (14, 15) and that the depressive scores of healthy and depressed subjects could be continuously estimated from local time-series statistics of physical activity (16). Cho et al. showed the possibility of constructing a predictive model of mood level among patients with mood disorders by applying machine learning to wearable data, such as sleep, activity, and heart rate (17). Sano et al. showed that the application of machine learning to physiological and behavioral data acquired by wearable sensors and mobile phones could be used to classify students’ self-reported mental health statuses (18). However, the sample sizes or observation periods of these studies were insufficient, and they have not reached the point where they look beyond symptoms and pathological evaluations to detect disease onset. In particular, the construction of a highly accurate disease onset predictive model that applies machine learning to large-scale data has not yet been achieved.

In addition to wearable data, blood- and urine-sample data collected during regular medical examinations can also be used as biomarker candidates for mental illness. Multiple blood metabolites have been shown to be associated with depression severity, with tryptophan concentration being particularly useful in classifying depressed individuals (19, 20). However, the construction of a highly accurate mental illness onset event predictive model that combines daily behavioral data relating to sleep or physical activities acquired by wearable devices with biochemical data based on blood and urine samples regularly acquired during medical examinations has not been achieved. This is primarily because big data that link medical examination and wearable data did not exist at the time.

The objective of this study was to build a model that can objectively predict the possibility of onset of mental illness from wearable and medical examination data. We have been engaged in the construction of a database that links insurance claims and medical examination data from society-managed health insurance (21), and wearable data linked to personal health records (PHRs), provided by information and communications technology (ICT) services for society-managed health insurance members. The data specifically include sleep, physical activity, and heart rate metrics measured using a consumer-grade wearable Fitbit device. In this study, we built a machine learning predictive model for mental illness onset that was trained using past insurance claims data. This study is meaningful in that the construction of a predictive model of disease onset based on the objective monitoring of human daily life and health status can facilitate the detection and prevention of the onset of mental illness and contribute to resolving widespread mental health issues in society.

METHODS

Study Design and Dataset

We conducted analyses of a database constructed by JMDC Inc. (21), which includes information on more than seven million society-managed health insurance members in Japan (insured individuals and dependents). This was comprised of our subject base, which is on large-scale and includes insurance claims (e.g., outpatient, hospitalization, and dispensary) and medical examination results. Notably, it also includes a trove of wearable data.

The subjects in this study were chosen from those satisfying the following conditions:

- Society-managed health insurance member;
- Individual whose wearable data are linked to PHRs [service name: Pep Up (https://pepup.life)], which is an ICT service by JMDC Inc. for society-managed health insurance members;
- Individual having a history of continuously linking wearable data for more than 3 months (15+ days per month); and
- Individual undergoing regular medical examinations.

The survey period ran from August 2016 to January 2020. The data were anonymized. Written informed consent was waived owing to the retrospective nature of this study while the subjects agreed to have their data analyzed when linking their data to PHRs. The ethical review of this study was performed by the Ethics Committee of the Research Institute of Healthcare Data Science (RIHDS). Date of approval: October 26, 2020; Approval number: RI2020013. The procedures were carried out in accordance with the approved guidelines.

Outcome Definition

Mental illness differs by individual in its duration from onset to hospitalization. However, as taken from the insurance claim data, the start of the onset was unclear. Hence, we based the onset period on the time of the first mental illness consultation. Relevant insurance claim information included the following qualifying criteria:

- Administration of hypnotics, anxiolytics, or antidepressants: prescription of medicine corresponding to Anatomical Therapeutic Chemical (ATC) Classification System codes.

...
(i.e., N05B, N05C, and N06A) as defined by the European Pharmaceutical Market Research Association (EphMRA);

- Psychiatric visits: calculation of medical practices corresponding to the Quick Reference Table of Medical Score classification codes I002 and I002-2.

Because hypnotics and anxiolytics are effective treatments for mental illness, they are frequently prescribed as a temporarily treatment for surgical hospitalization. Thus, to strictly determine that the administration was connected to a mental illness, we chose the initiation of both “administration of hypnotics, anxiolytics, or antidepressants” and “psychiatric visits” as the qualifying outcome. Henceforth, this combined outcome is referred to as “mental illness onset”. Although this study did not target a specific type of mental illness, this definition approximately narrows the range of the targets to people with depression or anxiety symptoms.

**Exclusion Criteria**

Potential subjects who met the following conditions were excluded from the subject data of this study:

- Individuals with mental illness onset prior to linking wearable data, during discontinued linking of data, or within 3 months after starting data linking; and
- Individuals with different months of meeting the qualifying outcomes listed above.

**Wearable Data**

The wearable data used in this study were acquired from Fitbit wearable devices from Fitbit Inc. The device model differed among subjects, but high inter-device reliability was reported (22). The data were systematically acquired through the Fitbit application programming interface (API) and consolidated into a single-format database.

The information for sleep-related items collected for each sleep session included isMainSleep, timeInBed, minutesAsleep, minutesToFallAsleep, minutesAfterWakeup, deepMinutes, remMinutes, lightMinutes, wakeMinutes, and startTime data items. isMainSleep is a Boolean that determines whether the given sleep event is main or not, e.g., long- or short-term sleep. timeInBed is the time spent in bed; minutesAsleep is the total sleep time excluding awakening time; minutesToFallAsleep is the time until falling asleep; minutesAfterWakeup is the time until getting out of bed after waking up; remMinutes, lightMinutes, deepMinutes, and wakeMinutes are the times spent in each sleep stage [i.e., rapid eye movement (REM), light non-REM (NREM), deep NREM, and wakefulness, respectively]; and startTime is the bedtime.

The information for activity-related items collected for each day included steps, veryActiveMinutes, fairlyActiveMinutes, and lightlyActiveMinutes. The physical activity times according to activity intensity were assigned based on the heart rate measured by the Fitbit. The heart rate information was acquired each day from RestingHeartRate.

**Medical Examination Data**

The medical examination data used in this study were acquired from annual medical examinations, which are required by law to be carried out by businesses affiliated with society-managed health insurers. The measurement items included body mass index (BMI), blood pressure (systolic and diastolic), blood-test items [i.e., triglyceride (TG), high-density lipoprotein (HDL) cholesterol, low-density lipoprotein (LDL) cholesterol, aspartate aminotransferase (AST), alanine aminotransferase (ALT), gamma glutamyl transferase (GT), fasting blood sugar (FBS), and hemoglobin A1c (HbA1c)], urine-test items (i.e., urinary sugar and uric protein qualitative), and interview items related to smoking, drinking, and exercise habits.

**Data Preprocessing**

The predictive model in this study predicted the presence of the onset of mental illness within 1 month using three continuous months of Fitbit and medical examination data closest to (i.e., within or prior to) the period. When a single subject had multiple time series that could be used as training data, all of them were used. Only the time series at the time of mental illness onset, however, was used for individuals who actually had mental illness onset.

The features used to build the predictive model are shown in Table 1. The wearable data were collected each month, with monthly features for 3 months being created. These features were labeled “[featureName]n,“ where n = 1, 2, or 3 for the n-th continuous month of data. The sleep data automatically acquired from the Fitbit were used to calculate social jet lag, sleep regularity index (SRI), and chronotype, which express sleep habits that are closely associated with mental health. Social jet lag is an index that indicates circadian misalignment and is calculated as the absolute difference between the mid-point of the sleep hours on weekdays and weekends (23). A larger social jet lag is known to exacerbate various health risks (24, 25). SRI quantifies the regularity of daily sleep–wake rhythms and is calculated as the concordance rate of sleep–wake occurrences in 24-h units (26, 27). Decreased SRI has been associated with stress, depression, poorer academic performance, and increased illness risk. Chronotype indicates morningness–eveningness preference and is often measured using questionnaires (28, 29). It is also associated with mental illness (e.g., depression) (30, 31). In this study, the chronotype calculated from wearable devices was used as an index to calculate corrected mid-sleep time on weekends for those with sleep debt corrected by its magnitude (32).

The monthly features of wearable data are described as “[featureName]1–3,” corresponding to the 1st—3rd month of the continuous 3-month data-linked period [e.g., SRI 1 is the SRI of the first month (earliest) of the data-linked period, and SRI 3 is the SRI of the last month (latest) of the data-linked period].

**Machine Learning and Validation**

The XGBoost binary classification model, which is known to exhibit superior predictive performance using table data, was used as the machine learning algorithm (33). XGBoost is
TABLE 1 | Features used in building the predictive model.

| Feature name [heat rate] | Unit | Description |
|--------------------------|------|-------------|
| RHR | bpm | monthly average of daily RestingHeartRate |

| Feature name [activity] | Unit | Description |
|-------------------------|------|-------------|
| veryActiveMinutes | min | monthly average of daily veryActiveMinutes |
| fairlyActiveMinutes | min | monthly average of daily fairlyActiveMinutes |
| lightlyActiveMinutes | min | monthly average of daily lightlyActiveMinutes |
| Steps | steps | monthly average of daily steps measured |
| log_count.active | days | monthly number of days for activity data linkage |

| Feature name [sleep] | Unit | Description |
|----------------------|------|-------------|
| notMainSleep.minutes | min | monthly average of daily determined to be “isMainSleep=False” |
| notMainSleep.counts | times | monthly counts of sleep determined to be “isMainSleep=False” |
| log_count.sleep | days | monthly number of days for activity data linkage |
| timeInBed | min | monthly average of daily bedtime |
| minutesAsleep | min | monthly average of daily minutesAsleep |
| minutesToFallAsleep | min | monthly average of daily minutesToFallAsleep |
| minutesAfterWakeUp | min | monthly average of daily minutesAfterWakeUp |
| deepMinutes | min | monthly average of daily deepMinutes |
| remMinutes | min | monthly average of daily remMinutes |
| lightMinutes | min | monthly average of daily lightMinutes |
| wakeMinutes | min | monthly average of daily wakeMinutes |
| startTime.sleep | [hhmms] | monthly average of daily bedtime |
| SC_lag | min | monthly social jetlag |
| chronotype | [hhmms] | monthly chronotype |
| SRI | - | monthly Sleep Regularity Index |

| Health Examination Data | | |
|-------------------------|------|-------------|
| BMI | kg/m² | Body Mass Index |
| SBP | mmHg | Systolic Blood Pressure |
| DBP | mmHg | Diastolic Blood Pressure |
| TG | mg/dl | Triglyceride |
| HDL | mg/dl | HDL cholesterol |
| LDL | mg/dl | LDL cholesterol |
| AST | U/l | Aspartate aminotransferase |
| ALT | U/l | Alanine aminotransferase |
| GT | U/l | Gamma glutamyl transferase |
| FBS | mg/dl | Fasting blood sugar |
| HBA1C | % (NGSP) | Hba1c |
| US | - | Urinary sugar |
| UP | - | Urinary protein qualitative |

| Health examination data [questionnaire] | | |
|--------------------------|------|-------------|
| SMOKE | - | Do you habitually smoke? |

(Continued)
FIGURE 1 | Overview of the predictive model construction and validation. (A) The predictive model was built using the XGBoost algorithm. The inputs to the model were 3-month’s continuous wearable data and those of the medical examinations closest to (i.e., within or prior to) that period; the output was the presence or absence of mental illness over the following month, which was defined based on insurance claims data. (B) Predictive performance was evaluated using a repeated stratified 10-fold cross validation (CV). Because the dataset included different time series from a single person, partition division was conducted to avoid including the same person in the model training and testing data in each fold (group CV).
the evaluation index was ensured by conducting CV 10 times by changing the cutting method of the fold, and the average AUC was used as the predictive performance of the model constructed in this study (i.e., repeated CV).

**Statistical Analyses**
The intervention range and methods of preventing mental illness onset using the predictive model were investigated by comparing the density estimation curve of the onset probability that was output by the predictive model between individuals with and without onset. A specific threshold was set for the onset probability that was output by the prediction model based on the receiver-operating-characteristic (ROC) and density estimation curves, and the statistical significance of the onset probabilities of the two groups was verified using Fisher’s exact test. The subject records were created by merging the verification data of each fold in 10-fold CV.

The contribution of each feature to the outcome was confirmed using feature importance (gain), which indicates the contribution of each feature in the model-learning process. Because XGBoost is an algorithm that presumes interactions, the effect that fluctuations in each feature have on the predictive value cannot be unified and accurately shown. Therefore, we adopted a novel method of calculating the maximum value of the threshold and the fraction at which the threshold value becomes the mode value (cover) as well as the onset probability of the two groups divided by the threshold and the fraction at which the threshold value is the mode value (cover) at each branch of the model. The top-10 features of the feature importance (gain) generated when the XGBoost model was built were, in order of decreasing importance, notMainSleep.counts1, lightlyActiveMinutes3, SC_lag1, minutesAsleep1, remMinutes1, SRII, log_count.sleep2, DRINK, TG, and GT (Figure 6). When confirming the most important features of the wearable data from the perspective of the time series, sleep-related features were all in the first month, whereas the days having linked sleep data occurred during the second month, and the mild activity time occurred during the third month.

Table 3 shows the mode of the branches of the most important features and the percentage at which the threshold value became the mode value (cover) as well as the onset probabilities of the two groups. With regards to the onset probabilities in the grouping based on the threshold value in the training data, the onset probability was significantly higher in the group having a feature value higher than the threshold value than in the group having a feature value lower than the threshold for notMainSleep.counts1, lightlyActiveMinutes3, minutesAsleep1, TG, and GT. The onset probability was significantly higher in the group having a feature value lower than the threshold than in the group having a feature value higher than the threshold for remMinutes1, SRII, and log_count.sleep2. Regarding DRINK, subjects with a drinking habit (threshold < 2.5 for DRINK indicates a drinking habit; see Table 1) exhibited a higher tendency for onset probability than those without a drinking habit (p = 0.059). SC_lag1 did not exhibit any significant differences in onset probability between the two groups according to mode split value (p = 0.278). Furthermore, notMainSleep.counts1, lightlyActiveMinutes3, log_count.sleep2, and DRINK had the same threshold for all branches (cover of 100%), whereas SC_lag1, SRII, and TG had a relatively low cover of around 50%.

**RESULTS**
The dataset comprised of 37,856 time series and 4,612 subjects (see Figure 2 for the details of subject inclusion). Table 2 shows the demographic data for the subjects. Among the subjects, 24 individuals (24 time series) exhibited mental illness onset.

The average AUC from repeated CV was 0.712 [standard deviation (SD) = 0.02]. Figure 3 shows the ROC curve for a given CV (AUC = 0.711). The primary hyperparameters were max_depth = 2, min_child_weight = 16, colsample_bytree = 0.8, class_weight = 10, and nrounds = 137. Figure 4 shows the relationship between class-weight and predictive performance; changing the loss function to a weighted logloss improved the AUC from 0.63 (class_weight = 1) to 0.71 (class_weight = 10).

Figure 5 shows the density estimation curve of the onset probability output by the predictive model. When grouping this onset probability with a threshold value of 1%, the actual onset rate of the group having a predictive probability > 1% was 0.17% (n = 7,592, x = 13) and the actual onset rate of the group with predictive probability < 1% was 0.04% (n = 30,264, x = 11), and there was a significant difference in the onset rates of the two groups (p < 0.001).

The top-10 features of the feature importance (gain) generated when the XGBoost model was built were, in order of decreasing

**DISCUSSIONS**
**Summary of the Findings**
In this study, we used machine learning to build a predictive model that used sleep and activity data acquired from Fitbit wearable devices and medical examination records to establish the criteria leading to mental illness onset. The results showed that the robust evaluation index of the constructed predictive model had an AUC that exceeded 0.7 and exhibited a moderate level of predictive performance. The relative strength of the relationship between wearable data and mental illness when compared with medical examination data was shown from the feature importance obtained. The top-ranking features of the constructed model included wearable data, specifically those of sleep-related indices. Analysis of the results of the model built with machine learning suggested that sleep abnormalities, especially the destabilization of sleep rhythms, are associated with an increased illness onset probability and that sleep disturbances may be a predictor of mental illness onset. Furthermore, activity-related indices and medical examination data relating to alcohol consumption were included in the topmost features, and these were also suggested to be factors. Early intervention aimed at sleep stabilization has been shown to be potentially effective in the prevention of mental illness.
Possible Interpretations and Implications of the Findings
The model developed in this study had an AUC > 0.7 and was thought to have achieved a moderate level of accuracy. A predictive model was built using data acquired from a consumer-grade wearable device (i.e., Fitbit). A disadvantage of Fitbit is that its data processing algorithm has not been completely disclosed, and its details are unknown. However, the fact that
a good-accuracy model was built using the machine learning XGBoost method shows that actual societal applications can be expected (37). The accuracy of consumer-grade wearable devices has continually increased, and many validation studies have discussed methods for evaluating and utilizing the quality of such devices (38–51). Recent systematic reviews relating to the measurement accuracy of sleep and physical activity by Fitbit have shown that, although there are still areas for improvement, it is fundamentally accurate (52–55). The results of this study provide evidence that precision preventive behavioral medicine (56) can be achieved by longitudinally acquiring personal health-related data using wearable devices.

The dataset used in this study was unbalanced with very few positive examples, and the prediction performance could be greatly improved by using weighted logloss as the loss function. Top-ranked features in terms of feature importance included SRI and SC_lag, and it is thought that the generation of features that reflect sleep rhythms from acquired wearable data will contribute to improved accuracy. Meanwhile, because

| TABLE 2 | Demographic data of the subjects (N = 4,612). |
|----------|---------------------------------------------|
| Age (years), mean (SD) | 45.9 (9.1) |
| Gender (male), N (%) | 3,289 (71.3) |
| Body mass index (kg/m²), mean (SD) | 23.2 (3.3) |

N, number of subjects; SD, standard deviation.

**FIGURE 3** | Receiver-operating-characteristic (ROC) for merged validation data created by 10-fold cross validation. Area-under-the-curve (AUC) = 0.711. The point closest to the top left (0,1) was (0.23, 0.67), and the corresponding cut-off value was 0.9%.
there are many negative examples despite the high AUC, as shown at the point closest to the top left on the ROC curve in Figure 3, interventions need to be conducted on \( \sim 23\% \) of all data-linked individuals to cover 67% of those having mental illness onset, owing to the percentage of negative cases \( \geq 99\% \). As shown in Figure 5, the predicted probability of the model is at most a few percent per month, and even with a high threshold, many of the cases can be false positives. Although false positive cases are “healthy” in the sense that psychiatrists do not administer hypnotic, anxiolytic, or antidepressant drugs in the following month, the risk of mental health problems may be relatively high. Therefore, soft interventions based on the high rate of false positives are required in this way, and specific candidates for this include advice without concerns about side effects relating to sleep, eating habits, or exercise.

The top-seven features contributing to the prediction of onset of mental illness included information acquired by wearable devices, indicating the importance of daily life monitoring of health metrics. In particular, six out of the seven features are related to sleep and, as mentioned in many previous studies (57–61), these results suggest the possibility of a close association between sleep abnormality and mental illness onset. Because activity-related features were also ranked second in importance, and three items acquired from the medical examination data were also included in the top-10 items, physical activity and medical examination data were confirmed to contribute to the diagnosis of mental illness onset to a degree.
In this study, we not only applied machine learning to data in order to build a model having high predictive accuracy, but we also confirmed the directionality of the associations between features and illness onset probability by investigating the detailed data using the model (Table 3). The manner in which the relationship between each feature and the outcome was determined was inferred by the mode values of the branches, the percentage at which the threshold value became the mode value (cover), and the onset probabilities of two groups with a grouping based on those values. Features having a cover of over 80% where the onset probability significantly differed between groups ($p < 0.05$) (notMainSleep.counts1, lightlyActiveMinutes3, minutesAsleep1, TG, GT, remMinutes1, SRI1, and log_count.sleep2) had a monotonic relationship with the outcome, suggesting a strong correlation by them without the premise of interaction. There were no significant differences in the onset probability between groups with regard to DRINK ($p = 0.059$), but the cover was 100%, suggesting that interactions may strengthen its effect on the outcome. The feature importance of SC_lag1 was ranked third, but no significant differences in onset probability were seen between groups ($p = 0.278$), and the cover was relatively low at 41%. Hence, it is suggested that the optimal threshold value changed significantly depending on whether the effect on outcome was quadratic functional (high risk even if the value was too large or too small) or by interactions. When confirming branches where that of SC_lag1 was a mode split value (105), it can be seen that the SC_lag1 branch was configured after (log_count.sleep2 $\geq 30$) = TRUE for all cases. When grouping
was conducted based on whether $SC_{lag1}$ was above or below the mode split value for the records with $log\_count\.sleep2 \geq 30$ in the training data, the onset probability of the “less than” group was 0% (0 / 14,367), and the onset probability of the “more than” group was 0.27% (3 / 1,108), with significant differences in the onset probability between groups ($p < 0.001$).

Among the sleep-related features, the most important, notMainSleep.counts1, was evaluated as high risk when notMainSleep exceeded five occurrences in a month. Increases in notMainSleep suggested increases in short-term sleep or polyphasic sleep, such as napping, and it is speculated that the decreased quality of MainSleep (i.e., sleep disturbances) is a background factor in mental illness onset. These results are consistent with the results of a previous study in which the number of naps was included in the top-ranking features used to predict the mental health status of students (18). Increases in $SC_{lag1}$ were evaluated as high risk, and this is consistent with the results of previous studies that showed that increased social jet lag increased various health risks (24, 25). However, it should be noted that the relationship of $SC_{lag1}$ may change according to the interactions with other features. Increases in minutesAsleep1 were evaluated to be at high risk. The threshold value was 7 h, and these results may be counterintuitive. However, the long period regardless of sleep quality, represented by deepMinutes, may reflect the loss of the sleep–wake rhythm and longer periods spent lying down or increased physiological

![Graph showing top-10 features in feature importance (gain) of the XGBoost model built using all of the training data.](image-url)
need for long periods of sleep caused by psychosomatic burdens, thereby increasing the illness onset risk. A \( \text{remMinutes1} \) value less than \( \sim 75 \) min was evaluated as high risk. A previous study using a polysomnography showed that shortened REM sleep latency was associated with depression (62), whereas the disruption and shortening of REM sleep was associated with poor physical and mental health (63–67). Some Fitbit models have been reported to have favorable REM sleep estimation accuracies (46), but care should be taken with regards to the sleep-stage estimation accuracy from consumer-grade wearable devices. Smaller SRII was evaluated as high risk, and this is consistent with previous studies that stated that irregular daily sleep times led to poor mental and physical states (60, 68, 69). Unlike conventional evaluative indices, which focus only on the main sleep phase, SRI is attracting attention as an index that can evaluate the sleep–wake rhythm (71). Unlike conventional evaluative indices, which focus only on the main sleep phase, SRI is attracting attention as an index that can evaluate the sleep–wake rhythm (71). Unlike conventional evaluative indices, which focus only on the main sleep phase, SRI is attracting attention as an index that can evaluate the sleep–wake rhythm (71).

With regards to activity-related features, it was confirmed that \( \text{lightlyActiveMinutes3} \), which was the second-highest in terms of feature importance, was evaluated as high risk when exceeding 260 min. This threshold value was in the top 20% of the overall record. Mental illness and depressive symptoms have been associated with low activity and slow movement (14–16, 70), and it is thought that there is a possibility of increased light activity intensity in the form of decreased physical activity immediately before illness onset, alongside disruptions in sleep rhythms and decreased balance.

The 8th–10th-ranked features in terms of feature importance were medical examination data, which included drinking habits (\( \text{DRINK} \)), triglycerides (\( \text{TG} \)), and \( \gamma \)-GTP (\( \text{GT} \)). All features were related to dietary lifestyle, including alcohol consumption. Drinking habits and high values indicating liver dysfunction were evaluated as high risk, and the results showed a link between dietary lifestyle and mental illness. Meanwhile, the importance of each of these features was lower than that of the wearable device-related features, and the results once again suggested the importance of wearable device data for the predictive detection of mental illness.

Focusing on temporal (monthly) information of wearable device-related features in terms of feature importance, the highly ranked sleep-related features were those from the first month, whereas the number of days with linked sleep data included features from the second month in the top rankings. This suggests the background of sleep habit disturbances first occurring prior to mental illness onset, which then lead to behavioral changes (e.g., taking off the Fitbit when going to sleep). It is interesting that the information gleaned prior to the month of the psychiatric consultation was important. The existence of a critical slowing down is known to be an early warning signal for a phase transition within complex dynamical systems (71). Moreover, reports have indicated that there are larger mood fluctuations immediately prior to the onset or termination of depression (72), as well as increased circadian rhythm instability in animal models prior to the establishment of alcohol dependence (73). From the results obtained regarding the manifestation of sleep rhythm instability, we can conclude that it provides an early warning for mental illness onset. Therefore, interventions that promote sleep improvement, in which abnormalities appear at an early stage, are thought to be effective.

Finally, although this study constructed a generalized model that predicts mental illness onset, individual differences can be understandably high. Individual differences may arise from complex interactions among biological/genetic, psychological/behavioral, and social/environmental factors. Therefore, utilizing a longitudinal personal time-series model with multiple, continuous, and objective data and records for one person should be more promising to achieve precision psychiatry (9) as well as precision preventive behavioral medicine (56). Machine learning and AI approaches should be very powerful tools for this purpose.

### TABLE 3 | Mode values of branches in the most important features and significant differences in onset probability owing to related groups.

| Feature                      | Mode split | Cover | Probability Less than | Probability More than or equal | p-value |
|------------------------------|------------|-------|-----------------------|--------------------------------|---------|
| notMainSleep.counts1         | 5          | 100%  | (62/62)               | 0.02% (4/20,471)               | 0.16% (17/10,453) | <0.001  |
| lightlyActiveMinutes3        | 260        | 100%  | (20/20)               | 0.04% (13/30,008)              | 0.14% (11/7,848) | 0.005   |
| SC_tag1                      | 105        | 41%   | (7/17)                | 0.06% (20/34,264)              | 0.11% (4/3,556) | 0.278   |
| minutesAsleep1               | 419        | 92%   | (23/25)               | 0.06% (21/36,840)              | 0.30% (5/1,016) | 0.026   |
| remMinutes1                  | 75         | 88%   | (15/17)               | 0.15% (7/4,553)                | 0.03% (2/7,326) | 0.032   |
| SRI1                         | 0.92       | 54%   | (7/13)                | 0.09% (22/23,637)              | 0.01% (2/14,219) | 0.002   |
| log_count.sleep2             | 30         | 100%  | (14/14)               | 0.09% (21/22,381)              | 0.02% (3/15,475) | 0.006   |
| DRINK                        | 2.5        | 100%  | (9/9)                 | 0.08% (21/25,439)              | 0.02% (2/8,946) | 0.059   |
| TG                           | 608        | 54%   | (13/24)               | 0.06% (22/37,776)              | 2.50% (2/80)    | 0.001   |
| GT                           | 19         | 91%   | (10/11)               | 0.01% (1/10,431)               | 0.08% (23/27,425) | 0.010   |

*Mode split cover* is the percentage of the mode in all branches used by the applicable feature. "Probability" is the onset probability of being less than or more than the mode split value relative to the entire training data. Records missing a feature value were excluded. The "p-value" was calculated using Fisher’s exact test. GT, glutamyl transferase; SRI, sleep regularity index; TG, triglycercide.
Limitations
The results in this study are from a dataset having only a small number of people with mental illness (24). The feature importance shown in this study also cannot be said to be sufficiently universal. It is possible that features unique to these 24 people were captured, and that other important features were overlooked. The interactions between the features in this study may not have been sufficiently trained with just 24 positive examples, and it may not have fully captured these aspects. In fact, \( \text{max}\_\text{depth} = 2 \) was used for hyperparameter selection by grid search, and the predictive model that was built in this study does not incorporate complicated interactions. Efforts must be made to continuously expand the dataset and to re-verify it.

The mental illness onset defined within this study’s scope can be captured from health insurance claims. However, many mentally ill people do not go to the hospital, even with obvious symptoms such as insomnia and depression. Nevertheless, the threshold for psychiatric consultations is particularly high in Japan compared with other countries. Hence, using variables of “administration of hypnotics, anxiolytics, or antidepressants” and “psychiatric visits” in this study is thought to have only captured in relatively severe cases. Clinically, onset is synonymous with confirmed diagnosis. Thus, for this study, the insurance paradigm was acceptable. Mental illness generally shows high variance in symptoms as well as underlying causes. As most of the positive cases in this study are thought to be people with depression and anxiety symptoms, the cases can be used to construct disease-specific predictive models in future studies.

This study used values measured by a specific device (i.e., Fitbit); the measurement items may not provide sufficient accuracy or may have proprietary/unclear features. This would not be a problem if we were to restrict our scope to Fitbit wearers. However, when comparing these results to those obtained using other wearable devices, appliance differences must be taken into account. Although there are limits to the generalizability of the results in this study, they are compelling in that they demonstrate the possibility of predicting illness onset using wearable data.

The period of the wearable data analyzed had to be restricted to three continuous months to secure a reasonable amount of data. The closest annual medical examination data to the mental illness onset was used in the analyses, which means that the timing of the medical examination was not the same as that of the wearable data. However, medical examination data generally do not change abruptly in such a short period (~several months), which should support the validity of our study design.

Biases associated with wearable data linkages should also be noted. This was an observational study, and it is always an individual's choice to attach and use wearable devices for any reason. Therefore, there is the possibility that a person who feels physically and mentally fulfilled will continue to wear the device and link multiple time-series records. Conversely, those who feel mentally offset may discontinue wearing the device. Hence, the onset probability predicted by the model is likely lower than that actually found in society as a whole. Nevertheless, there is some value in considering the utilization of a model that focuses on the magnitude of predictive probability and conducting preventive interventions based on a given threshold value.

CONCLUSIONS
This study demonstrated the possibility of developing a machine learning model that predicts mental illness onset using wearable data collection items and extant medical examination data. The feature importance in the predictive model developed in this study suggests that metrics such as sleep and activity cycles may be more useful in predicting the onset of mental illness than blood-test data. Sleep disturbances were detected as symptoms 3 months prior to onset, and early stage intervention that focused on improving sleep showed the potential for effective prevention.

DATA AVAILABILITY STATEMENT
The datasets presented in this article are not readily available because the data were used under license for the current study. JMDC makes the database widely available, on a fee-paying basis, for use in surveys, research, and commercial purposes. Requests to access the datasets should be directed to JMDC’s website (https://www.jmdc.co.jp/en/bigdata/).

ETHICS STATEMENT
The studies involving human participants were reviewed and approved by Ethics Committee of the Research Institute of Healthcare Data Science. The Ethics Committee waived the requirement of written informed consent for participation.

AUTHOR CONTRIBUTIONS
TS and AK were involved in the study conception and analytical design. TS extracted the data and performed the data analysis. TS, HS, and AK contributed to data interpretation. TS and AK wrote the first draft and all authors provided input on and approved the final manuscript. All authors contributed to the article and approved the submitted version.

FUNDING
AK was supported by The University of Tokyo Excellent Young Researcher project and JST PRESTO Grant Number JPMJPR19J3, Japan.

ACKNOWLEDGMENTS
The authors thank Mr. Masayuki Asano and Mr. Tadashi Miyahara for their helpful discussions regarding the direction of the study.
37. Baron KG. Feeling validated yet? A scoping review of the use of consumer-targeted wearable and mobile technology to measure and improve sleep. *Sleep Med Rev.* (2018) 40:151–9. doi: 10.1016/j.smrv.2017.12.002

38. Depner CM, Cheng PC, Devine JK, Khosla S, de Zambotti M, Robinillard R, et al. Wearable technologies for developing sleep and circadian biomarkers: a summary of workshop discussions. *Sleep.* (2020) 43:2xs254. doi: 10.1093/sleep/psz254

39. Svensson T, Chung UD, Tokuno S, Nakamura M, Svensson AK. A validation study of a consumer wearable sleep tracker compared to a portable EEG system in naturalistic conditions. *J Psychosom Res.* (2019) 126:109822. doi: 10.1016/j.jpsychores.2019.109822

40. Bent B, Goldstein BA, Kibbe WA, Dunn JP. Investigating sources of inaccuracy in wearable optical heart rate sensors. *NPJ Dig. Med.* (2020) 3:18. doi: 10.1038/s41746-020-0226-6

41. Bliwise DL, Chappell C, Maidlisch L, Roitmann E, Burtea T. A multitrait, multimethod matrix approach for a consumer-grade wrist-worn watch measuring sleep duration and continuity. *Sleep.* (2021) 44:zsaa141. doi: 10.1093/sleep/zsaa141

42. de Zambotti M, Cellini N, Goldstone A, Cellini N, Goldstone A, Baker FC. Wearable sleep technology in clinical and research settings. *Med Sci Sports Exerc.* (2019) 51:1538–57. doi: 10.1249/MSS.0000000000001947

43. Walch O, Huang Y, Forger D, Goldstein C. Sleep stage prediction with raw acceleration and photoplethysmography heart rate data derived from a consumer wearable device. *Sleep.* (2019) 42:zs180. doi: 10.1093/sleep/zs180

44. Wright SP, Hall Brown TS, Collier SR, Sandberg K. How consumer physical activity monitors could transform human physiology research. *Am J Physiol Regul Integr Comp Physiol.* (2017) 312:R358–67. doi: 10.1152/ajpregu.00349.2016

45. Cook JD, Prairie ML, Plante DT. Utility of the Fitbit Flex to evaluate sleep in major depressive disorder: a comparison against polysomnography and wrist-worn actigraphy. *J Affect. Disord.* (2017) 217:299–305. doi: 10.1016/j.jad.2017.04.030

46. de Zambotti M, Goldstone A, Claudatos S, Colrain IM, Baker FC. A validation study of Fitbit Charge 2™ compared with polysomnography in adults. *Chronobiol Int.* (2018) 35:465–76. doi: 10.1080/07420528.2017.1413578

47. Kahawage P, Jamrakhyob R, Hamill K, de Zambotti M, Drummond SPA. Validity, potential clinical utility, and comparison of consumer and research-grade activity trackers in Insomnia Disorder I: In-lab validation against polysomnography. *J Sleep Res.* (2020) 29:e12931. doi: 10.1111/jsr.12931

48. Hamill K, Jumabhoy R, Kahawage P, de Zambotti M, Walters EM, Drummond S. Validity, potential clinical utility and comparison of a consumer activity tracker and a research-grade activity tracker in insomnia disorder II: outside the laboratory. *J Sleep Res.* (2020) 29:e12944. doi: 10.1111/jsr.12944

49. Roomkham S, Lovell D, Cheung J, Perrin D. Promises and challenges in the use of consumer-grade devices for sleep monitoring. *IEEE Rev Biomed Eng.* (2018) 11:53–67. doi: 10.1109/RBME.2018.2811735

50. Chinoy ED, Cuellar JA, Huwa KE, Jameson JT, Watson CH, Bessman SC, et al. Performance of seven consumer sleep-tracking devices compared with polysomnography. *Sleep.* (2021) 44:zsaa291. doi: 10.1093/sleep/zsaa291

51. Massoomi MR, Handberg EM. Increasing and evolving role of smart devices in modern medicine. *Eur Cardiol.* (2019) 14:181–6. doi: 10.15420/ecr.2019.02

52. Straiton N, Alharbi M, Bauman A, Neubeck L, Gullick J, Bhindi R, et al. The validity and reliability of consumer-grade activity trackers in older, community-dwelling adults: a systematic review. *Maturitas.* (2018) 112:85–93. doi: 10.1016/j.maturitas.2018.03.016

53. Fehan LM, Geldman J, Sayre EC, Park C, Ezzat AM, Yoo JY, et al. Accuracy of Fitbit devices: systematic review and narrative syntheses of quantitative data. *JMIR mHealth uHealth.* (2018) 6:e10527. doi: 10.2196/10527

54. Haghayegh S, Khoshnevis S, Smolensky MH, Diller KR, Castriotta RJ. Accuracy of wristband Fitbit models in assessing sleep: systematic review and meta-analysis. *J Med Internet Res.* (2019) 21:e16273. doi: 10.2196/16273

55. Fuller D, Colwell E, Low I, Orychcock K, Tobin MA, Simango R, et al. Reliability and validity of commercially available wearable devices for measuring steps, energy expenditure, and heart rate: systematic review. *JMIR mHealth uHealth.* (2020) 8:e18694. doi: 10.2196/18694

56. Saito et al. Machine Learning Mental Illness Prediction
Conflict of Interest: TS and HS were employees of JMDC Inc.

The remaining author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher’s Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Saito, Suzuki and Kishi. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.