Uncovering social states in healthy and clinical populations using digital phenotyping and Hidden Markov Models

Imogen E. Leaning[*a,b], imogen.leaning@donders.ru.nl
Andrea Costanzo[c], a.costanzo@rug.nl
Raj Jagesar[c], r.r.jagesar@rug.nl
Lianne M. Reus[d,e], l.reus@amsterdamumc.nl
Pieter Jelle Visser[d,f,g], pj.visser@maastrichtuniversity.nl
Martien J.H. Kas[c], m.j.h.kas@rug.nl
Christian Beckmann[a,b], christian.beckmann@donders.ru.nl
Henricus G. Ruhé[a,h], eric.ruhe@radboudumc.nl
Andre F. Marquand[a,b,i], andre.marquand@donders.ru.nl

[*]Corresponding author.

[a]Donders Institute for Brain, Cognition and Behaviour Radboud University Nijmegen, Nijmegen, the Netherlands
[b]Department for Cognitive Neuroscience, Radboud University Medical Center Nijmegen, Nijmegen, the Netherlands
[c]Groningen Institute for Evolutionary Life Sciences, University of Groningen, Groningen, the Netherlands
[d]Department of Neurology, Alzheimer Center, Amsterdam Neuroscience, Amsterdam UMC, Amsterdam, the Netherlands
[e]Center for Neurobehavioral Genetics, Semel Institute for Neuroscience and Human Behavior, David Geffen School of Medicine, University of California, Los Angeles, USA
[f]Department of Psychiatry & Neuropsychology, School for Mental Health and Neuroscience, Maastricht University, Maastricht, the Netherlands
[g]Department of Neurobiology, Care Sciences and Society, Division of Neurogeriatrics, Karolinska Institutet, Stockholm, Sweden
[h]Department of Psychiatry, Radboud University Medical Center Nijmegen, Nijmegen, the Netherlands
[i]Department of Neuroimaging, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, United Kingdom

Abstract

Brain related disorders are characterised by observable behavioural symptoms. Smartphones can passively collect objective behavioural data, avoiding recall bias. Despite promising clinical utility, analysing smartphone data is challenging as datasets often include a range of missingness-prone temporal features. Hidden Markov Models (HMMs) provide interpretable, lower-dimensional temporal representations of data, allowing missingness. We applied an HMM to an aggregate dataset of smartphone measures designed to assess social functioning in healthy controls (HCs) (n=247), participants with schizophrenia (n=18), Alzheimer's disease (AD) (n=26) and memory complaints (n=57). We selected a model with socially "active" and "inactive" states, generated hidden state sequences per participant and calculated their "dwell time", i.e. the percentage of time spent in the socially active state. We identified lower dwell times in AD versus HCs and higher dwell times related to increased social functioning questionnaire scores in HCs, finding the HMM to be a practical method for digital phenotyping analysis.

Introduction

Many psychiatric and neurological diseases exhibit observable behaviours that are indicative of the underlying condition. For example, social functioning is negatively impacted in a broad range of conditions, including schizophrenia (SZ), Major Depressive Disorder (MDD), anxiety disorders and Alzheimer's disease (AD) (e.g.[1-3]), often cumulating in social withdrawal. Social withdrawal, indicated by reduced social interaction[1], can be observed as people engage less with those around them. However, successfully measuring behavioural components such as social withdrawal is challenging, as reports of behaviour are subjective and susceptible to recall bias, with questionnaires often being burdensome to complete. There is therefore a need to develop practical, objective tools to monitor these symptoms, for example to predict or measure clinically relevant changes.

The field of digital phenotyping is developing to meet such a need. Digital phenotyping involves the development of behavioural or physiological markers calculated from digital measures (i.e. "digital phenotypes"). These measures avoid issues of recall bias as they are objective and can be acquired in real-time as participants go about their day, also meaning they have high ecological validity. A popular tool to collect digital phenotyping data is the smartphone. Given the highly common place of smartphones in society, the smartphone is convenient as it does not require participants to change their behaviour or routines; a monitoring application, for example "Behapp",[4] "Mood mirror"[5] or "RADAR-base pRMT",[6] can be installed on their own phone and run passively in the background to collect data, without user intervention.

Modern smartphones have a large number of sensors and functionalities, including various applications (apps), calling capabilities, WiFi, Global Positioning System (GPS), accelerometer and Bluetooth, which can be leveraged to model different aspects of behaviour (including social contacts, movement patterns and app usage (e.g. [7], [8])). Moreover, there are many ways in which these data can be processed. For example, the duration, circadian rhythm or statistical measures can be calculated (such as mean and standard deviation of a behaviour across time) or the occurrences of the behaviour counted.[9] This often leads to datasets with many features reflecting various different smartphone-measured behaviours. A major problem affecting digital phenotyping is that platforms are often prone to missing data due to the difficulties of real-world longitudinal data collection, leading to missing values across all or a subset of these features.[9]

This gives rise to multiple analytic challenges: processing the collected feature sets, often representing a wide range of seemingly distinct observed behaviours with potentially similar underlying causes, requires many model decisions. Appropriate methods are therefore needed to analyse this multi-faceted data containing missingness, in order to produce meaningful, lower-dimensional data representations. These representations may be more usable and informative about the underlying behavioural states of participants relative to the individual features. Models should also aim to be interpretable by not only researchers but also clinicians (and patients), to facilitate their use in clinical practice. A further property that would enable their use in this

context is that they can preserve the time domain, as one of the goals of smartphone digital phenotyping is to be able to make useful clinical predictions that can enable early intervention. Many digital phenotyping studies have focused on time-averaged features and analyses, and a shift towards more direct investigations of temporal dynamics is expected to improve clinical utility.[9]

Currently, digital phenotyping studies employ a broad range of approaches. For example, investigating associations between neuropsychiatric symptoms and summary measures (e.g. total number of places visited, mean duration of communication app usage),[10] clustering of digital phenotypes to investigate transdiagnostic symptom classification,[11] linear mixed effects models accounting for repeated measures of time-averaged features ([12-15]), multivariate anomaly detection to identify relapse in SZ[16] and joinpoint regression to identify changes in the trajectory of digital phenotypes (e.g. step count).[17]

In this study we propose the use of a Hidden Markov Model (HMM) (e.g.[18]) as a method to model digital phenotyping time series data. This provides several appealing features, namely that HMMs 1) can meaningfully combine different behavioural features, 2) reflect changes in behaviour over time, 3) provide readily interpretable summary statistics and 4) naturally accommodate missingness. HMMs provide interpretable, lower dimensional representations of the data using latent (i.e. hidden) states, where the observed time series channels are represented as a sequence of these hidden states. Each hidden state has associated "emission probabilities" indicating the probability that the state corresponds to the observed behaviours, allowing for informative behavioural states to be derived by representing more than one feature per state. Changes in behaviour through time are modelled via transitions between these hidden states. Importantly for digital phenotyping, HMMs contain intrinsic mechanisms for handling missing data. HMMs have been used in many applications for modelling behaviour, for example to model drinking patterns in people with an alcohol use disorder,[19] cocaine dependence,[20] sleep patterns represented in neuroimaging data,[21] and mobility data (e.g. [22], [23]).

While our approach is widely applicable to digital phenotyping time series, in this work we demonstrate its application to data collected using the Behapp monitoring application (www.behapp.com), which collects passive data related to app usage, calls, GPS, WiFi and overall phone usage, reflecting the periods the phone was unlocked. We applied an HMM to a combined dataset of phone usage and communication-related features from participants in the "Psychiatric Ratings using Intermediate Stratified Markers" (PRISM)[24] and Hersenonderzoek (HO)[10] studies, demonstrating how an HMM can successfully represent digital phenotyping time series. The model was initially trained on a set of HCs with low missingness to provide a high-quality dataset for training, and then applied to HCs with higher missingness, and participants with AD, SZ and healthy participants with memory complaints ("subjective cognitive complaints"; SCC), to investigate generalisability. Hidden state sequences were generated for these participants, and we then calculated a digital phenotype derived from the HMM for each participant, namely the "dwell time". The dwell time provides the percentage of time the participant spent in a hidden state. This digital phenotype was then linked to clinical measures

including diagnostic group and social functioning, demonstrating the clinical value of this approach.

Results

Sample statistics

An overview of our approach is shown in Figure 1. This study utilised data from participants in the PRISM and HO datasets, which jointly contained 247 HCs, 18 participants with SZ, 26 with AD and 57 participants with SCC. Participants with AD and HCs were present in both datasets, whereas participants with SZ were provided by PRISM and participants with SCC were provided by HO.

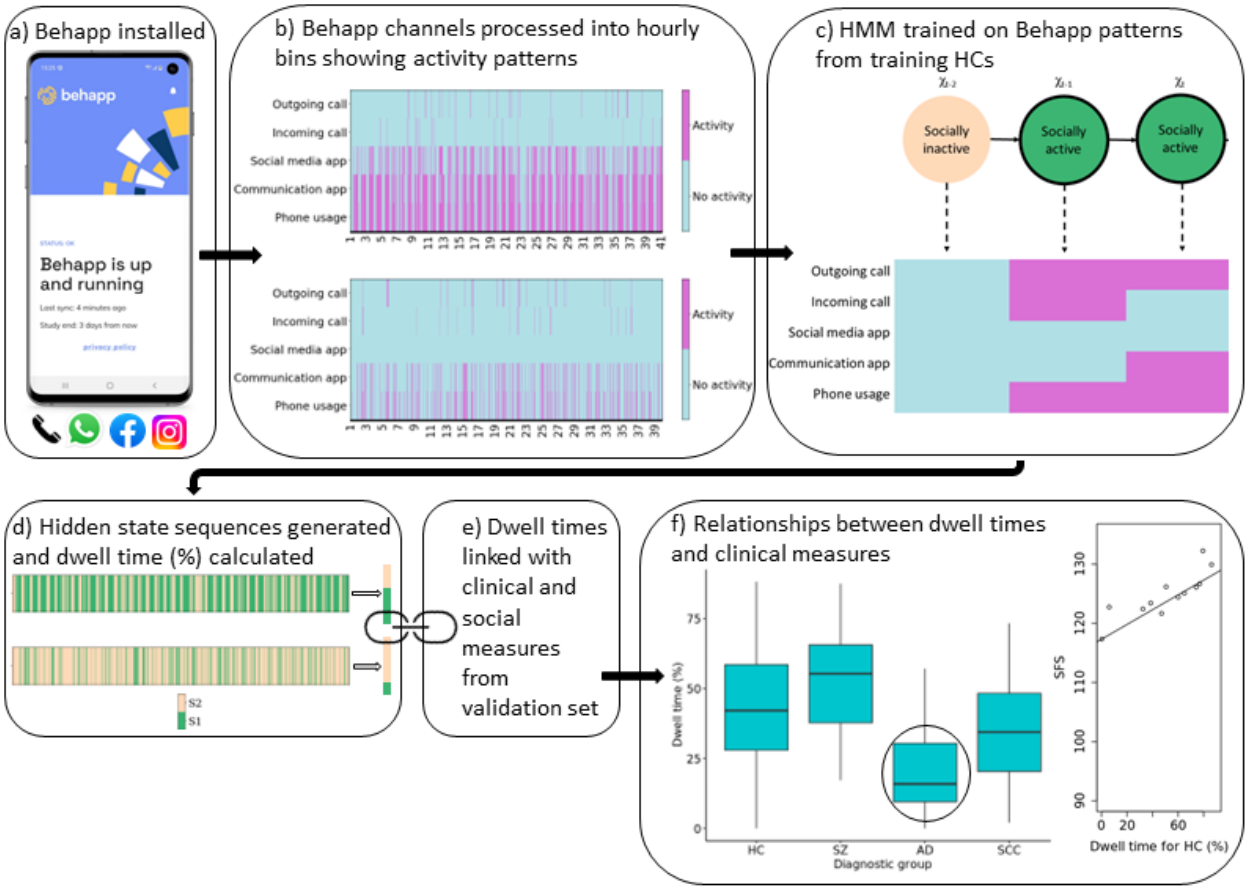

Figure 1: Flowchart highlighting the main processing and modelling steps involved in the HMM approach. Panel f) shows our main clinical findings: lower socially active dwell time in AD versus HCs, and a positive relationship between socially active dwell time and social functioning in HCs. HC: healthy control, SZ: schizophrenia, AD: Alzheimer's disease, SCC: Healthy/subjective cognitive complaints.

In the PRISM and HO datasets, HCs were age matched to the diagnostic groups, with the PRISM sample being matched to both SZ and AD and the HO sample age-matched only to AD. After aggregation of datasets, this results in a bimodal age distribution. More specifically, due to the expected differences in age between participants with SZ and AD, the HCs are on average older than participants with SZ and younger than those with AD. However, note that the

difference in age between the diagnostic groups is a consequence of aggregating multiple samples. From the age histograms presented in Figure 2, it is clear that the HC group spans the full range of each diagnostic group, and we also performed additional sensitivity analyses with matched diagnostic groups to confirm group comparison findings. Training set and overall validation set age distributions are shown in supplemental Figure S1.

PRISM data was collected across sites in the Netherlands and Spain, whilst HO data was collected solely in the Netherlands. PRISM recorded participant race, with nearly all participants identifying themselves as white, whereas HO did not report participant race. The demographics of the HCs, split by training versus validation set assignment, are provided in supplemental Table S1.

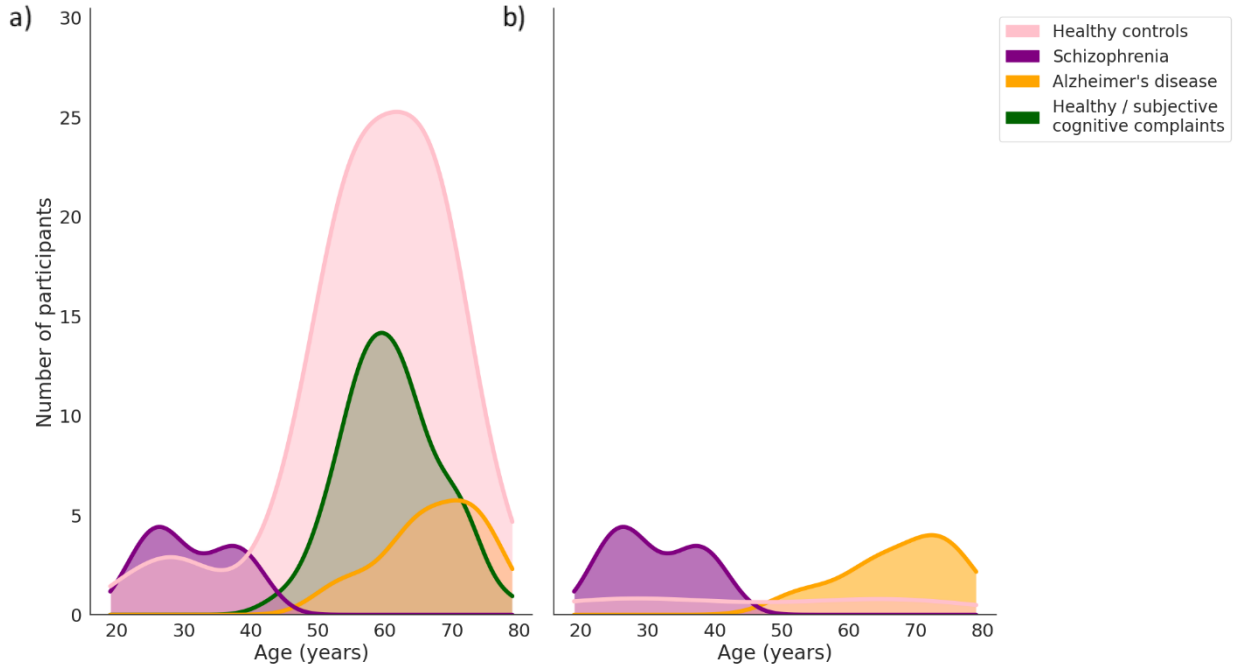

Figure 2: a) Distribution of ages for all validation participants, b) distribution of ages for validation participants with social measures. Plotted using kernel density estimation.

Hidden Markov Model derivation and interpretation

When training the HMM, the number of hidden states used by the model must be chosen. During our hyperparameter selection, we evaluated two- and three-state models, which both converged, however it was seen that one of the states in the three-state model comprised a very small percentage of the state sequences for the training segments (<2%) and was viewed as redundant (a four-state model was consequently not investigated). Two was therefore chosen as an appropriate number of states for the model, and the two-state model with the highest model likelihood was used in the subsequent analyses.

The emission probabilities of the states generated by the two-state model are shown in Figure 3. Using these emission probabilities to interpret the hidden states, it is evident that they represent socially active and socially inactive states. That is, the first state (S1) corresponds to phone usage with a very high probability that communication apps are also being used by the participant. There is a smaller probability of social media usage and outgoing/incoming phone calls. This state also includes a low probability of missing data. Due to the use of communication methods in this state, such as calls and app usage, this hidden state is referred to as the "socially active" hidden state. The second state (S2) corresponds to a much smaller probability of phone usage, with the probability of all other channels near zero, and is referred to as the "socially inactive" hidden state. We show a demonstrative example of how the hidden states correspond to the observed channels in Figure 4, illustrating different observed channel configurations that can correspond to each of the hidden states.

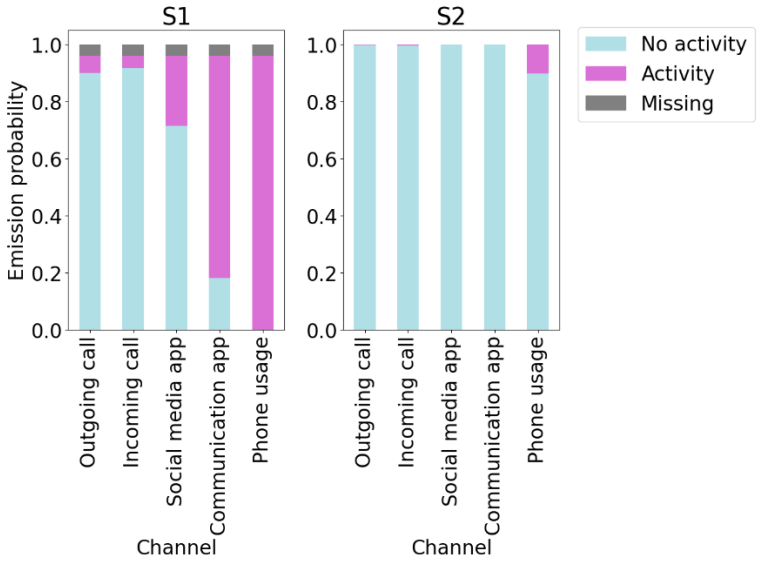

Figure 3: Emission probabilities of the selected two-state model. S1: State 1, S2: State 2.

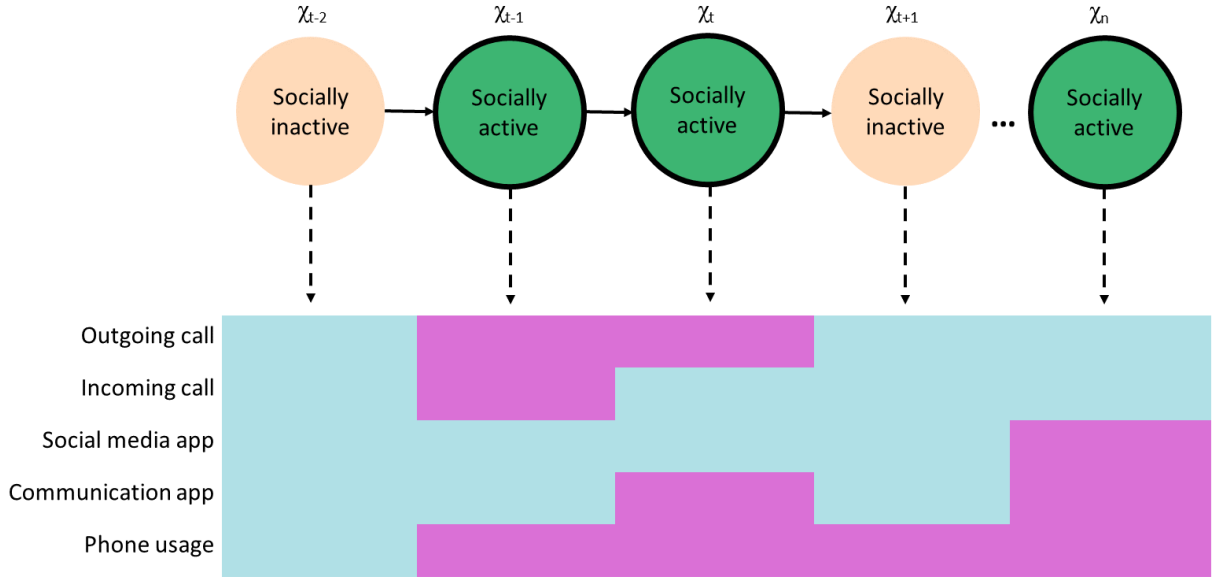

Figure 4: Examples of which behaviours may correspond to the hidden states. Socially active state: various social behaviours are displayed, including calls and app usage; socially inactive state: no phone usage, or phone usage without corresponding social behaviours.

After model training, the hidden state sequence corresponding to each participant's time series was generated. The dwell time for each validation participant can then be calculated from the hidden state sequence, with missing data in the validation set removed, and compared to clinical scores and diagnostic group. We chose to drop the missing portions from these time series before hidden state sequence generation, otherwise a participant with, for example, half of their time series missing would show all of this missing period as being socially active (see Figure 3). As the selected model only contains two states and the dwell time (i.e. the proportion of time spent in each state) is a percentage value, only the dwell times corresponding to one of the states needed to be investigated. We therefore focus on the dwell times from the "socially active" state, and so further reference to "dwell time" derived from the HMM solely refers to dwell times in the socially active state.

An example of one participant's hidden state sequence alongside the input sequence is shown in Figure 5, and an example of another participant can be seen in Figure 6. It is immediately apparent that the subject shown in Figure 5 spends considerably more time in the socially active state relative to the subject shown in Figure 6. It can be seen that the participants in both Figure 5 and Figure 6 oscillate quite frequently between the socially active and inactive states, which is not surprising due to expected diurnal variation (e.g. [25]). More clearly, higher social activity during the daytime and lower social activity during night-time can be seen in Figure 7. Additionally, the probability of starting a hidden state sequence in the socially active and inactive states were 0.246 and 0.754 respectively, showing that it is more probable to begin

the time series in the socially inactive state. This is to be expected as all of the time series began at midnight, so many participants would have been asleep.

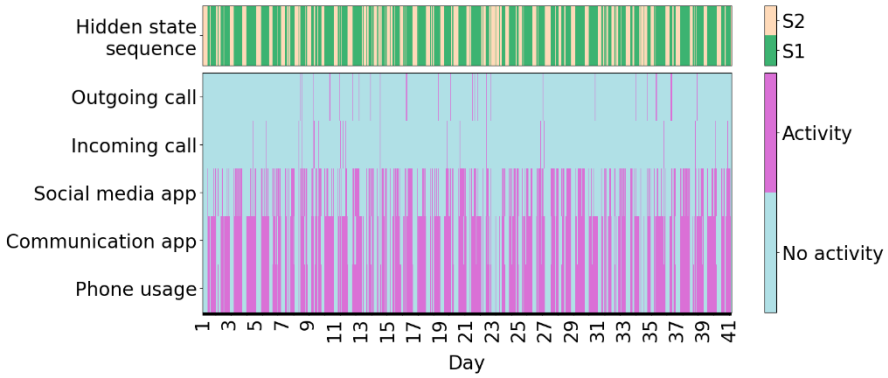

Figure 5: The observed time series composed of hourly bins (bottom five rows) of a participant compared with their corresponding predicted hidden state sequence (top row); S1: State 1 (socially active state), S2: State 2 (socially inactive state).

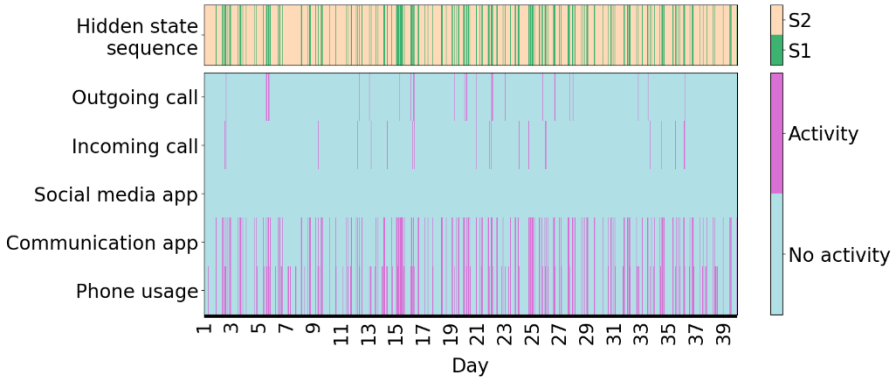

Figure 6: The observed time series composed of hourly bins (bottom five rows) of another participant compared with their corresponding predicted hidden state sequence (top row); S1: State 1 (socially active state), S2: State 2 (socially inactive state).

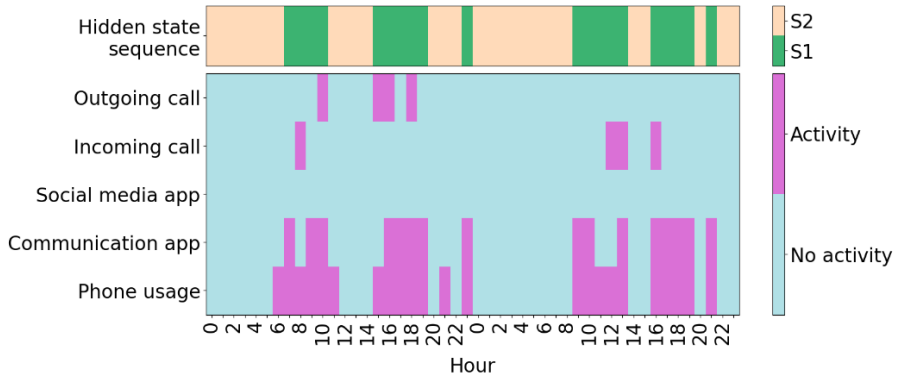

Figure 7: An example of a two-day period of a participant's time series; S1: State 1 (socially active state), S2: State 2 (socially inactive state); 0: midnight.

Measures of social functioning and loneliness

For validation purposes, we make use of a measure of social functioning for each participant in the PRISM dataset, namely the Social Functioning Scale (SFS)[26] (see Figures S2 and S3 in the supplementary materials for score distributions). We therefore investigated possible relationships between social functioning and socially active dwell times for the various available participant groups in the validation set (i.e. HCs, participants with SZ or AD). The number of participants in each group is small, so we consider our results to be preliminary indicators of possible relationships between the HMM-derived digital phenotypes and social functioning.

Linear regression models were run to investigate possible relationships between SFS scores and dwell times, with age included as an additional predictor in the models. Separate models were run for each of the diagnostic groups, with one model run for all groups combined (Table 2). FDR corrected *p*-values (considering four tests) are presented with results considered significant at $p < .05$. A significant positive relationship between social functioning and dwell times was found for the HCs (FDR corrected *p*-value = 0.0041), with every one percent increase in dwell time corresponding to a 0.1248 increase in SFS score, but no significant relationship was found for the other diagnostic groups. For the overall validation group, an effect of age was found (FDR corrected *p*-value = 0.0136), however this effect was driven by the lower SFS scores of the SZ group.

A measure of loneliness[27] was also provided for the PRISM participants, however no significant relationship between loneliness and dwell times was found. The results from these linear regression models are presented in supplemental Table S2, as well as histograms of the distribution of loneliness scores (Figures S4 and S5).

Diagnostic group

A multinomial logistic regression model was run to investigate differences in socially active dwell time between the different diagnostic groups and the HC group in the validation set (i.e. the reference category) (see Figure 8). Age was again included as an additional predictor in the model, and FDR corrected *p*-values (considering three tests) are presented to provide an indicator

of significance at $p < .05$ (Table 3). Dwell time was found to be a significant predictor of AD relative to HCs (FDR corrected p-value = 0.0002); participants with AD generally showed lower dwell times (i.e. spending less time in the socially active state) relative to HCs (odds ratio = 0.9455). No significant relationship of dwell time on SZ or SCC group was found relative to HCs. Due to the broad age range of HCs, sensitivity analyses of age were carried out for each diagnostic group (supplemental Table S3), with a subsample of HCs age-matched to each respective diagnostic group, with the AD result remaining significant (FDR corrected p-value = 0.0003).

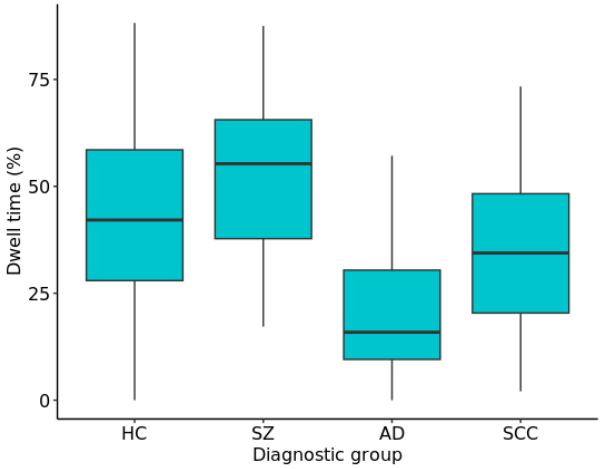

Figure 8: A box plot of the dwell times for the different diagnostic groups (HC: healthy control, SZ: schizophrenia, AD: Alzheimer's disease, SCC: Healthy/subjective cognitive complaints); there is a significant difference between the HC and AD groups.

Further clinical measures

For participants with AD and the HCs in the PRISM dataset, Mini-Mental State Examination (MMSE)[28] scores, measuring cognitive impairment, were provided. Whilst across validation participants an effect of dwell time was found on MMSE (FDR corrected p-value = 0.0157), no significant relationship was found within diagnostic groups, so it is likely that this significant value is capturing the group differences of HC versus AD, rather than a specific effect of dwell time on MMSE (see Table 4). No significant effect of age was found on MMSE (age-related model results are provided in Table S4 in the supplementary materials, as well as score distributions in Figures S6 and S7).

The PRISM dataset also provided Positive and Negative Syndrome Scale (PANSS)[29] scores for participants with SZ, however no significant relationships between any of the PANSS scores (positive, negative, general psychopathology, composite and total) and dwell time were found. The results from these linear regression models are presented in Table S5 in the supplementary materials, as well as histograms of the distribution of PANSS scores per subscale (Figure S8).

Discussion

A central aim of digital phenotyping is to develop objective measures that can be used to monitor clinically-relevant behaviours and symptom changes. In this study we proposed a method for deriving meaningful, interpretable digital phenotypes using the HMM, a time series model that can accommodate missingness. We applied this model to general phone usage and communication smartphone measures, calculating the socially active dwell time phenotyped by the HMM. Our smartphone measures were collected passively, reducing burden on participants, and we protected participant privacy by abstracting app measures to descriptive levels, without collecting content. We investigated the association of the socially active dwell time with various social and clinical measures, including diagnostic group and a questionnaire on social functioning (SFS). We found that a two-state HMM, that switches between socially active and socially inactive states, could suitably represent the participants' five-channel smartphone time series. We observed a significant difference in the HMM-derived "socially active" dwell times between HCs and participants with AD, with participants with AD exhibiting lower dwell times. This difference was robust to age sensitivity analysis. A significant relationship between dwell times and social functioning was also identified for HCs, with higher dwell times corresponding to higher social functioning.

The HMM has several strengths: it uses lower-dimensional hidden states to represent the various observed behaviours, which can be easily interpreted for each state using the emission probabilities (Figure 3). The socially active state could be interpreted as being linked to observed communication-related behaviours, whilst the socially inactive state reflected a lack of these behaviours, such as other kinds of, or no, phone usage. Transitions between these hidden states were indicative of behavioural changes throughout time, for example daily behavioural patterns (Figure 7). The HMM can also handle missing data points during model estimation, incorporating a probability of missingness into one of the hidden states (see Figure 3). Hidden states may allow for some individual behaviours to be represented as comparable behaviours. For example, Figure 6 shows a time series with no social media usage, whereas Figure 5 shows highly recurrent social media usage, and both of these participants can have their respective behaviours represented using the socially active state despite individual differences in what social activity may mean for each participant. This type of modelling approach can therefore allow for a certain amount of flexibility in the behaviours of the participants, dependent on the number of hidden states used in the model.

A summary measure of the HMM, the socially active dwell time, was calculated per validation participant so that a model-derived digital phenotype could be compared to clinical and social measures. The observed difference in dwell time between participants with AD and HCs, with AD dwell times lower than HCs, is consistent with the understanding that AD is associated with impaired social functioning,[1] and demonstrates a potential objective measure of this difference. A positive relationship between social functioning and dwell time was identified for HCs. Given the large influence of communication-related behaviours contributing to the socially active state, this relationship may in part be reflecting higher scores in the

communication-related questions in the SFS. Regarding cognitive impairment, an overall significant relationship between dwell times and MMSE was found, but not when separate models were run for HCs and participants with AD. This difference is likely reflecting group differences in dwell time between HCs and AD.

Differences in dwell time relative to HCs were not observed for SCC participants or participants with SZ. These results may be unsurprising as by definition SCC participants are very similar to HCs, with the difference in inclusion criteria being that SCC participants experience memory complaints. Similarly, the participants with SZ did, for the most part, exhibit quite low symptom severity. The number of participants with SZ was also small. Whilst the PRISM study only placed exclusion criteria on positive symptoms (to exclude psychosis), the negative symptoms in the sample did not turn out to be very severe either, and overall most participants could be classified as "mildly ill" based on their total PANSS score.[30] This is indicative of a selection of less affected patients. The mild PANSS scores, as well as low loneliness scores, may also contribute to the absence of an identified relationship between these scales and dwell time. Significant relationships between social functioning and dwell time for participants with AD and SZ were also not observed. It is possible that participants with AD and SZ may overestimate their social functioning,[31] which could be reflected in their self-report SFS scores. This may complicate any possible relationship between this social functioning measure and dwell time for these groups. A further interesting factor that could affect these relationships is the impact of different symptom profiles on dwell time.

To expand upon the current work, the HMM method could be applied in a larger population of SZ participants exhibiting broader symptom severity and different symptom profiles. Given the reluctance of many people with acute psychotic symptoms to being monitored, it may be necessary to monitor participants for a longer period of time, beginning with low symptom severity at study enrolment, to allow for more fluctuations in symptom severity to be observed (e.g. [16]). The HMM method can also be applied to other disorders, including Major Depressive Disorder (to be included in PRISM 2). A wider range of smartphone channels can also be included in the HMM, for example calls could be encoded to reflect the variation in who is called/is calling each hour. With a larger number of input channels, the derived hidden states could reflect more specific behavioural states. The optimal number of hidden states may then be driven both by the number of input channels, and the underlying behavioural states of the participants themselves. With a higher order model, the hidden states' emission probabilities would not necessarily correspond to distinct single behaviours; for example with the inclusion of GPS channels, there could be two hidden states that correspond to time spent at home, with one state also reflecting communication activities and the other reflecting no communication.

For the current analysis, each hidden state sequence was generated per participant, but dwell time comparisons were only made between groups. To shift towards individual predictions (for example predicting symptom scores or relapse along the time series), the dwell time for windows of the sequence, or potentially the sequence likelihood, could be extracted and changes

along the time series evaluated. This would also maintain the time component of the analysis; our current analysis uses a time series model but then compares a summary HMM measure to clinical measures. For clinical applications, the eventual goal would be to be able to make individual predictions along the time series.

To improve the management of missing data, there are several more avenues that can be explored. Data is often expected to be missing due to technical difficulties, but it is also possible that data can be missing due to user behaviour, for example if the user switches the phone off, turns on flight mode or deletes the app from their phone. Future studies could consider recording these specific behaviours (which would currently be more feasible with Android phones, rather than iOS), to provide a better indicator of data missing due to technical difficulties versus user behaviour. With regards to managing missing data specifically in the context of HMMs, with higher dimensional input data it may be appropriate to allow for separate "missing states". If missingness occurs in certain channels but not others, it is possible for this to be reflected in the hidden states (e.g. a state corresponding to missing GPS data but present phone usage).

Due to high rates of missingness, we made four main decisions to handle missing data: 1) to focus model training on high data availability time series, 2) to use a model that can accommodate missing data, 3) to exclude GPS channels from the current time series analysis due to low levels of data availability affecting these specific channels and 4) to exclude missing timepoints from the validation time series before hidden state sequence generation. Whilst we view decisions 1) and 2) as useful strategies for managing missing data, decision 3), and to a lesser degree decision 4), were unfortunate consequences that in future studies should be avoided with improved data collection. The datasets used in this study were collected with early versions of Behapp, and throughout data collection no indicator of missingness was known. Indicators of missing data were developed retrospectively using WiFi and GPS sampling frequencies to assist analyses of these time series. Incoming data monitoring has now been improved in more recent Behapp versions, as well as the overall data collection process. Researchers using Behapp can therefore now track data collection as it is ongoing, and take action if sustained periods of data are missing. This could involve contacting participants to ensure they have not accidentally disabled desired functionalities for sustained periods.

For interpretation purposes, we have named the two hidden states as "socially active" and "socially inactive". However a person could, of course, be socially active offline without using their phone. For example, a person may be socialising with friends at home without using their phone. We therefore acknowledge limits to our naming convention, and recommend caution when interpreting hidden states. Other sensors could be used to give an indicator of other people in the participant's vicinity, such as Bluetooth[32], but passive smartphone data will nevertheless remain somewhat of a proxy for social activity. In a similar vein, we used the App Store classification to group apps, but participants may use the apps for purposes other than this classification (e.g. some people use Instagram for communication, and less so for social media). Whilst in our two-state model these discrepancies would be inconsequential, with a larger number of hidden states these discrepancies could potentially lead to misleading interpretations

of a person's behaviour. In clinical application, the patient's behaviours could be discussed with the clinician at the beginning of Behapp usage to assist in understanding and interpreting their personal digital phenotypes.

Smartphone-based digital phenotyping is a promising tool for monitoring and predicting mental health outcomes. However, methods are needed for managing this multi-faceted time series smartphone data. We proposed the use of an HMM to model digital phenotyping time series, as this method can 1) combine different behavioural features, 2) reflect temporal behavioural changes 3) be easily interpreted and 4) manage missingness. We developed a two-state model that represented various smartphone channels as "socially active" and "socially inactive" states, and calculated the socially active dwell time for each participant's time series. We identified a significant difference between HC and AD dwell times, with AD dwell times lower than HCs, showing how this HMM-derived digital phenotype may be a useful measure to indicate differences in social functioning. We also observed a significant positive relationship between dwell time and social functioning for HCs, which could reflect the increase in communication behaviours in the socially active state and their connection with social functioning. The HMM is an interpretable method to model behaviour based on digital phenotyping data and with further development provides an appealing approach for making clinical predictions of symptom changes and relapse across a range of neuropsychiatric diseases.

Method

Participants

This study utilised data from participants in PRISM and HO. We chose to combine these datasets to increase sample size, and due to the overlap in Alzheimer's populations and consequentially, similarly age-matched HCs.

PRISM

The PRISM study aims to investigate social withdrawal in two brain disorders, SZ and probable AD ([24,33]). Participants with AD, SZ, and age- and gender- matched HCs were recruited across centres in Spain (Hospital General Universitario Gregorio Marañón and Hospital Universitario de La Princesa, in Madrid) and the Netherlands (University Medical Center Utrecht, Leiden University Medical Center and Amsterdam UMC, location VUmc).

Participants with SZ were required to be within the age range of 18-45 years (inclusive), and to have a DSM-IV (Diagnostic and Statistical Manual of Mental Disorders) diagnosis of SZ confirmed by the Mini-International Neuropsychiatric Interview (MINI). Participants were required to have experienced at least one psychotic episode, to have had a maximum disease duration of 10 years since diagnosis, and for any antipsychotic medication dosage to have been stable for a minimum of 8 weeks. As PRISM aimed to investigate social withdrawal linked with negative symptoms (and not as a consequence of other sources such as psychosis), participants with SZ were excluded if they rated highly for positive symptoms (≥22 on the positive symptom factor of the 7-item Positive and Negative Syndrome Scale (PANSS)[29]). Participants with AD were required to be within the age range of 50-80 years, to meet the classification of "Probable AD" based on the National Institute on Aging and the Alzheimer's Association (NIAAA) criteria, and to have a Mini-Mental State Examination (MMSE)[28] score of 20-26. For both participants with SZ and AD, it was required that participants were not socially withdrawn due to other reasons such as their external circumstances, a comorbid medical disorder or disability.

HCs were recruited in the age ranges of 18-45 and 50-80, and were required to have an approximately average MMSE score according to their age and years of education. Participants were excluded if they met the criteria for an Axis-I psychiatric disorder (assessed by the MINI), or a neurological disease associated with cognitive impairment. For further details of inclusion/exclusion criteria for all participant groups see the PRISM study overview.[24]

In addition to Behapp data collection, measures of clinical and social functioning were acquired. The self-report Social Functioning Scale (SFS)[26] and the De Jong Gierveld Loneliness and Affiliation Scale[27] were administered to all participants, the MMSE was administered to HCs and participants with AD, and the PANSS was administered to participants with SZ.

Hersenonderzoek

Participants with probable AD, "Healthy/subjective cognitive complaints" (SCC) participants and age-matched HCs were recruited across the Netherlands by The Dutch Brain Research Registry (Hersenonderzoek.nl), providing demographics and health-related information online via the Hersenonderzoek.nl platform.[10] Participants indicated the presence of probable AD. To

classify participants as SCC or HC respectively, participants indicated the absence of neurological or psychiatric diseases with or without memory complaints. The minimum age for inclusion was 45 years.

Ethical approval and informed consent

PRISM was approved by the Ethics Review Board of each of the five participating centres in Spain and the Netherlands, and participants were deemed by the researcher and caregivers to be sufficiently competent to participate. Approval for HO was provided by the Ethical Review Board VU University Medical Centre. All participants provided informed consent before participation commenced.

Behapp acquisition

The smartphone application "Behapp" (www.behapp.com) was installed on participants' smartphones. Behapp passively collected smartphone-usage data for a period of 42 days without storing any content of messages and calls, in compliance with the European Privacy Regulation.[34] The classification of each app used by participants was gathered from the Google Play Store, so that apps could be grouped by type, including social media and communication apps. During the time of data collection (PRISM: August 2017 – May 2019; HO: March 2018 – January 2020), Behapp was only available on Android smartphones, and so PRISM participants who did not have their own Android smartphone were supplied with one for the duration of study participation. However, this was not done for HO participants in accordance with the study design, and only two PRISM participants used a study-provided phone. For each activity (e.g. use of an app), the respective start and end timestamps were stored.

Preprocessing

Smartphone channels

Phone usage was split into five categories, referred to as "channels": social media app usage, communication app usage, incoming calls, outgoing calls, and overall phone usage. GPS channels were also available. Since many of these measures are sparsely sampled, each channel was aggregated into hourly bins, and the percentage of each hour for which each activity was carried out was calculated. For example, a participant may spend 100% of an hour using their phone, 50% on social media, 40% on communication apps, 0% making/receiving calls and 10% using another functionality such as Google Maps. Even with the temporal resampling, many of these phenotypes have highly zero-inflated distributions (see Figure S9 in the supplementary materials), which can be difficult to handle natively. Therefore, for each hourly timepoint, these percentages were grouped into discrete bins instead of continuous percentages: binary bins reflecting either no or some activity carried out in the hour (0% activity; >0-100% activity).

Two measures were developed to identify whether data had been correctly collected by Behapp for each hour, which capitalise on the sampling frequency of the location and other data sources like WiFi data (which are both independent of personal phone usage); as this frequency is expected to be greater than once per hour, a frequency lower than one sample per hour in the location data indicates missing location data, and a frequency lower than one sample per hour in

all types of data (including WiFi) indicates that overall data was not being collected successfully. Therefore one of these measures reflected overall data availability, and the other measure was specific to GPS data availability. The distributions for these measures are provided in supplemental figures S10-S13. These measures were required so that we could differentiate between values that were zero because a participant was not using their phone, and values that were zero because data was not successfully collected. Due to low GPS data availability acquired using the version of Behapp used in these studies, it was decided not to include the GPS channels in the current analysis. Therefore any missingness that occurs in the included channels occurs across all channels at the same timepoints (i.e. it is not possible to have data missing at a timepoint in, for example, only the social media channel and not the other channels).

To account for any changes in behaviour that may have arisen from study onboarding (i.e., participant attending assessments at study location), the first day of each participant's Behapp data was excluded. As a consequence, all time series began at midnight. If the overall data availability measure indicated missing data, then the channels were marked as "NaN". Since missing data are handled natively by the HMM implementation we employed,[18] as explained below, no missing data imputation was carried out on the data.

Division into training and validation sets based on missing data and diagnostic group

Participants were split into training and validation sets, with the training set used to train the model and the validation set used to investigate relationships between HMM-derived digital phenotypes and clinical measures. All participants with SZ, AD or SCC were assigned to the validation set (as well as a subset of withheld HCs), so that the HMM could be trained on HCs, akin to training on a reference category. To ensure that the HMM was trained on high quality data (i.e. time series with low levels of missingness), HCs meeting an overall data availability criterion of at least 90% of timepoints available across their time series were assigned to the training set. No minimum requirement was set for Behapp participation length, so that shorter time series that did not have missingness issues during data collection were still included. We randomly selected 15 of these high data availability HCs and retained them in the validation set, to allow for some amount of data availability matching between HCs in the training and validation sets, also increasing the number of HCs in the validation set with social and clinical scale measures available. The distributions of time series lengths for training and validation participants are provided in Figures S14 and S15 in the supplementary materials, and distributions of data availability are provided in Figures S10-S13.

Overview of Hidden Markov Model

The HMM models the observed smartphone data channels using a smaller number of hidden states, where each hidden state corresponds to probable values in these observed channels. Through time the participant then switches between different hidden states. The HMM model was implemented and fitted using the R package "depmixS4".[18] During model training, the expectation-maximization algorithm is used to maximise the expected joint log-likelihood of the model parameters. The depmixS4 package allows for missing values in the dataset, which means

that missing values are effectively omitted from the calculation of the log-likelihood. Each response variable (i.e. observed channel) was modelled using a multinomial distribution with an identity link function. This is the preferred multinomial link function in depmixS4 when no covariates are present, due to its computational speed. As all of the input channels were binned into binary bins to manage the zero-inflation, this resulted in a binomial distribution for each response variable.

We investigated a range in the number of hidden states, and for each hidden state number ran the model ten times with different random seed initialisations. As the input data included a total of five channels, the possible number of hidden states used by the HMM ranged from two to four. Due to the small range of possible hidden states, this hyperparameter was not formally optimised, but selected based on observations regarding the composition of the hidden state sequences corresponding to each model order. After the number of hidden states was chosen, the HMM model using the seed which provided the highest model likelihood for this number of states was selected and used in subsequent analyses.

We then applied the trained HMM to the validation dataset and generated the hidden state sequences corresponding to these participants' time series using the Viterbi algorithm. Note that the hidden state sequence is equal in length to the observed time series. In our implementation, applying the trained HMM to the validation dataset does not involve retraining the model; this step is simply required due to the different time series lengths of the validation participants compared to the lengths in the training set.

Hidden Markov Model measures

Various probabilities reflecting each of the hidden states are learnt during model training, which can be used to describe the model and to understand what behaviours each of the hidden states are associated with. This includes emission, starting and transition probabilities:

Emission probability: The emission probability for each state refers to the probability that a hidden state corresponds to given values in each of the observed channels, and can therefore be used to interpret what observed behaviours each hidden state represents. A state may correspond to activity in some observed behavioural channels and not others, and this can be identified with the emission probability.

Starting probability: The starting probability indicates the probability of beginning the sequence in each hidden state. If a time series often begins with the same observed values, then the hidden state corresponding to these values will have a high starting probability.

Transition probability: The transition probability gives the probability of switching into another hidden state from each state (or the probability of staying in the same state). For example, for behaviours with long durations, the transition probability of staying in the associated hidden state may be high relative to the probability of transitioning to a non-related hidden state.

Additionally, other measures can be calculated from the hidden state sequence itself. In this study we focus on a measure referred to as the "dwell time":

Dwell time: The dwell time per hidden state, also known as fractional occupancy, gives the percentage of time during which a state was occupied. This can be calculated for any desired level of granularity, for example, for all participants together, for each participant, or for a specific time period. In this study the dwell time was calculated per participant in the validation set, as we had a single value from each scale available per participant, i.e. no repeated measures. As the validation set contained a range of data availability, any missing data timepoints were dropped from the time series before hidden state sequence generation, so that the calculation of dwell time only reflected the available data.

As we used a two-state model in this study, we concentrated solely on the dwell time in the socially active state, and do not refer to a socially inactive dwell time in the analyses.

Comparison of dwell time to social and clinical measures

The dwell times were compared to two social measures using linear regression models: social functioning (SFS)[26] and loneliness[27] (available for participants in the PRISM study). For each of these scales, separate models were run for each of the diagnostic groups and one model was run for the combined groups. Dwell times were then compared between the different diagnostic groups and HCs (available for participants in both PRISM and HO) using multinomial logistic regression. Linear regression models were also run for clinical scales reflecting cognitive impairment (MMSE) and schizophrenia symptoms (PANSS) (available for the AD (and HCs) and SZ participants in the PRISM study respectively). In the case of PANSS scores, separate models were run for the total score and the subscores (positive, negative, general psychopathology and composite).

For all regression models age was included as a predictor, and for the logistic regression, sensitivity analyses of age were also carried out for each diagnostic group, due to the broad age range in HCs as a consequence of age-matching to both SZ and AD. For the SZ sensitivity analysis, the maximum SZ participant age was used as the maximum cut-off age for HCs, and for AD and SCC each respective minimum participant age was used as the minimum cut-off age for HCs. Binomial logistic regression models were then run for each diagnostic group compared to improved age-matched HCs.

Data availability

The datasets analysed during the current study are available from the corresponding author on reasonable request.

Code availability

All scripts used in the analyses are available at https://github.com/predictive-clinical-neuroscience/HMM_Digital_Phenotyping

Acknowledgements

This study was funded by the European Research Council (consolidator grant 101001118). The Dutch Brain Research Registry (Hersenonderzoek.nl) is supported by ZonMw-Memorabel (project no 73305095003), Alzheimer Nederland, Amsterdam Neuroscience, and Hersenstichting (Dutch Brain Foundation). The PRISM project has received funding from the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No 115916. This Joint Undertaking receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA. This study reflects only the authors' view and the European Commission is not responsible for any use that may be made of the information it contains.

Author contributions

IEL: conceptualisation, formal analysis, methodology, software, visualisation, writing – original draft
AC: data curation, investigation, software, writing – review & editing
RJ: data curation, software, writing – review & editing
LMR: data curation, writing – review & editing
PJV: data curation, writing – review & editing
MJHK: conceptualisation, funding acquisition, project administration, writing – review & editing
CB: conceptualisation, supervision, writing – review & editing
HGR: conceptualisation, supervision, writing – review & editing
AFM: conceptualisation, funding acquisition, methodology, supervision, writing – review & editing

Competing interests

Christian Beckmann is a director of SBGNeuro. Henricus Ruhé received grants from the Hersenstichting, ZonMw, the Dutch Ministry of Health and an unrestricted educational grant from Janssen. In addition he received speaking fees from Lundbeck, Janssen, Benecke and Prelum; all outside the current work.

References

1 Van der Wee, N. J. *et al.* Working definitions, subjective and objective assessments and experimental paradigms in a study exploring social withdrawal in schizophrenia and Alzheimer's disease. *Neuroscience & Biobehavioral Reviews* **97**, 38-46 (2019).

2 Porcelli, S. *et al.* Social brain, social dysfunction and social withdrawal. *Neuroscience & Biobehavioral Reviews* **97**, 10-33 (2019).

3 Saris, I. M. J., Aghajani, M., Van Der Werff, S. J. A., Van Der Wee, N. J. A. & Penninx, B. W. J. H. Social functioning in patients with depressive and anxiety disorders. *Acta Psychiatrica Scandinavica* **136**, 352-361 (2017).

4 Jagesar, R. R., Vorstman, J. A. & Kas, M. J. Requirements and operational guidelines for secure and sustainable digital phenotyping: Design and development study. *Journal of Medical Internet Research* **23** (2021).

5 Bai, R. *et al.* Tracking and monitoring mood stability of patients with major depressive disorder by machine learning models using passive digital data: prospective naturalistic multicenter study. *JMIR mHealth and uHealth* **9**, e24365 (2021).

6 Ranjan, Y. *et al.* RADAR-base: open source mobile health platform for collecting, monitoring, and analyzing data using sensors, wearables, and mobile devices. *JMIR mHealth and uHealth* **7**, e11734 (2019).

7 Eskes, P., Spruit, M., Brinkkemper, S., Vorstman, J. & Kas, M. J. The sociability score: App-based social profiling from a healthcare perspective. *Computers in Human Behavior* **59**, 39-48 (2016).

8 Jongs, N. *et al.* A framework for assessing neuropsychiatric phenotypes by using smartphone-based location data. *Translational psychiatry* **10**, 211 (2020).

9 Leaning, I. E. *et al.* From smartphone data to clinically relevant predictions: A systematic review of digital phenotyping methods in depression. *Neuroscience & Biobehavioral Reviews*, 105541 (2024).

10 Muurling, M. *et al.* Assessment of Social Behavior Using a Passive Monitoring App in Cognitively Normal and Cognitively Impaired Older Adults: Observational Study. *JMIR Aging* **5** (2022).

11 Kas, M. J. H. *et al.* Digital behavioural signatures reveal trans-diagnostic clusters of Schizophrenia and Alzheimer's disease patients. *European Neuropsychopharmacology* **78**, 3-12 (2024).

12 Pellegrini, A. M. *et al.* Estimating longitudinal depressive symptoms from smartphone data in a transdiagnostic cohort. *Brain and Behavior* **12**, e02077 (2022).

13 Tønning, M. L., Faurholt-Jepsen, M., Frost, M. & Kessing, L. V. Mood and activity measured using smartphones in unipolar depressive disorder. *Frontiers in Psychiatry* **12**, 701360 (2021).

14 Faurholt-Jepsen, M. *et al.* Differences in mobility patterns according to machine learning models in patients with bipolar disorder and patients with unipolar disorder. *Journal of Affective Disorders* **306**, 246-253 (2022).

15 Sun, S. *et al.* Challenges in Using mHealth Data From Smartphones and Wearable Devices to Predict Depression Symptom Severity: Retrospective Analysis. *Journal of medical Internet research* **25**, e45233 (2023).

16 Cohen, A. *et al.* Relapse prediction in schizophrenia with smartphone digital phenotyping during COVID-19: a prospective, three-site, two-country, longitudinal study. *Schizophrenia* **9**, 6 (2023).

17 Fujino, Y., Tokuda, F. & Fujimoto, S. Decreased step count prior to the first visit for MDD treatment: a retrospective, observational, longitudinal cohort study of continuously measured walking activity obtained from smartphones. *Frontiers in public health* **11**, 1190464 (2023).

18 Visser, I. & Speekenbrink, M. depmixS4: an R package for hidden Markov models. *Journal of statistical Software* **36**, 1-21 (2010).

19 Shirley, K. E. *Hidden Markov models for alcoholism treatment trial data*. (University of Pennsylvania, 2007).

20 DeSantis, S. M. & Bandyopadhyay, D. Hidden Markov models for zero-inflated Poisson counts with an application to substance use. *Statistics in medicine* **30**, 1678-1694 (2011).

21 Stevner, A. B. A. *et al.* Discovery of key whole-brain transitions and dynamics during human wakefulness and non-REM sleep. *Nature communications* **10**, 1035 (2019).

22 Witayangkurn, A., Horanont, T., Sekimoto, Y. & Shibasaki, R. in *Proceedings of the 2013 ACM conference on Pervasive and ubiquitous computing adjunct publication.* 1219-1228.

23 Baratchi, M., Meratnia, N., Havinga, P. J., Skidmore, A. K. & Toxopeus, B. A. in *Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing.* 401-412.

24 Bilderbeck, A. C. *et al.* Overview of the clinical implementation of a study exploring social withdrawal in patients with schizophrenia and Alzheimer's disease. *Neuroscience & Biobehavioral Reviews* **97**, 87-93 (2019).

25 Vesel, C. *et al.* Effects of mood and aging on keystroke dynamics metadata and their diurnal patterns in a large open-science sample: A BiAffect iOS study. *Journal of the American Medical Informatics Association* **27**, 1007-1018 (2020).

26 Birchwood, M., Smith, J. O., Cochrane, R., Wetton, S. & Copestake, S. O. N. J. A. The social functioning scale the development and validation of a new scale of social adjustment for use in family intervention programmes with schizophrenic patients. *The British Journal of Psychiatry* **157**, 853-859 (1990).

27 de Jong-Gierveld, J. Developing and testing a model of loneliness. *Journal of personality and social psychology* **53**, 119 (1987).

28 Folstein, M. F., Folstein, S. E. & McHugh, P. R. "Mini-mental state": a practical method for grading the cognitive state of patients for the clinician. *Journal of psychiatric research* **12**, 189-198 (1975).

29 Kay, S. R., Fiszbein, A. & Opler, L. A. The positive and negative syndrome scale (PANSS) for schizophrenia. *Schizophrenia bulletin* **13**, 261-276 (1987).

30 Leucht, S. *et al.* What does the PANSS mean? *Schizophrenia Research* **79**, 231-238 (2005).

31 Jongs, N. *et al.* Effect of disease related biases on the subjective assessment of social functioning in Alzheimer's disease and schizophrenia patients. *Journal of Psychiatric Research* **145**, 302-308 (2022).

32 Zhang, Y. *et al.* Predicting depressive symptom severity through individuals' nearby bluetooth device count data collected by mobile phones: preliminary longitudinal study. *JMIR mHealth and uHealth* **9**, e29840 (2021).

33 Kas, M. J. *et al.* A quantitative approach to neuropsychiatry: the why and the how. *Neuroscience & Biobehavioral Reviews* **97**, 3-9 (2019).

34 Mulder, T., Jagesar, R. R., Klingenberg, A. M., Bonnici, J. P. M. & Kas, M. J. New European privacy regulation: Assessing the impact for digital medicine innovations. *European Psychiatry* **54**, 57-58 (2018).

Tables

Table 1: Demographics of each of the diagnostic groups.

Diagnostic group	Number	Age (Mean ± std)	Gender	Dataset	Country	Education years (Mean ± std)
Healthy control	247	59 ± 13	F=140; M=107	PRISM=28; HO=219	NL=234; ES=13	6 ± 4
Schizophrenia	18	31* ± 6	F=7; M=11	PRISM=18; HO=0	NL=12; ES=6	15 ± 3
Alzheimer's disease	26	67* ± 7	F=10; M=16	PRISM=19; HO=7	NL=18; ES=8	13 ± 7
Healthy/subjective cognitive complaints	57	61 ± 7	F=36; M=21	PRISM=0; HO=57	NL=57	5 ± 2

**Statistically significant difference in age from HCs. std: standard deviation; F: female, M: male; NL: the Netherlands, ES: Spain.*

Table 2: Results from linear regression models predicting SFS score from dwell time; each row corresponds to a different model for different group-based splits of the validation set.

Group	Coefficient	Standard error	t value	*p*-value	FDR corrected *p*-value
All validation participants (n=49)	0.1148	0.0683	1.6793	0.0999	0.3995
Healthy control (from validation set) (n=12)	0.1248	0.0262	4.7647	0.0010	0.0041*
Schizophrenia (n=18)	0.1291	0.1078	1.1977	0.2496	0.9985
Alzheimer's disease (n=19)	-0.2676	0.1049	-2.5500	0.0214	0.0856

Table 3: Results from a multinomial logistic regression model predicting diagnostic group (versus healthy controls (n=156)) using dwell time (age was also included as a predictor).

Group	Coefficient	Standard error	Odds	Wald statistic	*p*-value	FDR corrected *p*-value
Schizophrenia (n=18)	-0.0107	0.0164	0.9894	-0.6544	0.5129	1.0000
Alzheimer's disease (n=26)	-0.0560	0.0142	0.9455	-3.9474	0.0001*	0.0002*
Healthy/subjective cognitive complaints (n=57)	-0.0165	0.0081	0.9836	-2.0347	0.0419*	0.1256

Table 4: Results from linear regression models predicting MMSE score from dwell time; each row corresponds to a different model for different group-based splits of the validation set.

Group	Estimate	Standard error	t value	*p*-value	FDR corrected *p*-value
All validation participants (n=31)	0.0610	0.0206	2.9696	0.0061	0.0182*
Healthy control (from validation set) (n=12)	0.0139	0.0107	1.3008	0.2256	0.6769
Alzheimer's disease (n=19)	0.0603	0.0339	1.7780	0.0944	0.2832