Did you miss me? Making the most of digital phenotyping data by imputing missingness with point process models

Authors:
Imogen E. Leaning [1,2]
Andrea Costanzo [3]
Raj Jagesar [3]
Loran Knol [1,2,4,5]
Sarah Tjeerdsma [6]
Anna Tyborowska [1,7]
Nessa Ikani [8]
Lianne M Reus [9,10]
Pieter Jelle Visser [9,11,12]
Martien J.H. Kas [3]
Christian F. Beckmann [1,2]
Henricus G. Ruhé [1,6]
Andre F. Marquand [1,2,13]

[1] Donders Institute for Brain, Cognition and Behaviour Radboud University Nijmegen, Nijmegen, the Netherlands
[2] Department for Medical Neuroscience, Radboud University Medical Center Nijmegen, Nijmegen, the Netherlands
[3] Groningen Institute for Evolutionary Life Sciences, University of Groningen, Groningen, the Netherlands
[4] Department of Psychiatry, University of Illinois at Chicago, Chicago, IL, USA
[5] Department of Biomedical Engineering, University of Illinois at Chicago, Chicago, IL, USA
[6] Department of Psychiatry, Radboud University Medical Center Nijmegen, Nijmegen, the Netherlands
[7] Behavioural Science Institute, Radboud University Nijmegen, Nijmegen, the Netherlands
[8] Department of Developmental Psychology, Tilburg School of Social and Behavioral Sciences, Tilburg University, Tilburg, the Netherlands
[9] Department of Neurology, Alzheimer Center, Amsterdam Neuroscience, Amsterdam UMC, Amsterdam, the Netherlands
[10] Amsterdam Neuroscience, Neurodegeneration, Amsterdam UMC, Amsterdam, the Netherlands
[11] Department of Psychiatry & Neuropsychology, School for Mental Health and Neuroscience, Maastricht University, Maastricht, the Netherlands
[12] Department of Neurobiology, Care Sciences and Society, Division of Neurogeriatrics, Karolinska Institutet, Stockholm, Sweden
[13] Department of Neuroimaging, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, United Kingdom

Abstract

Digital phenotyping has broad clinical potential, providing low-burden objective measures of behaviour as individuals go about their day. However, progress in making clinical inferences from these data is severely challenged by the common occurrence of missing data. We investigated non-homogeneous Poisson Point Process Models (PPPMs) as a method for imputing missing temporal digital phenotyping data, considering smartphone-based activities as 'points'. We assessed the inclusion of time-varying covariates ('hour' and 'day') in personalised PPPMs. We used participants from SMARD (n=26) for a ground truth evaluation, and PRISM (n=65) and Hersenonderzoek (n=283) for a replication analysis involving Hidden Markov Models, evaluating the effectiveness of PPPMs for imputation and their influence on downstream analysis. In the ground truth evaluation, PPPMs using one-hot encoded hour provided the highest out-of-sample likelihood. We successfully replicated findings from our prior work using this method, demonstrating that PPPMs using covariates are a promising imputation tool for digital phenotyping.

Keywords: passive sensing, imputation, mHealth, smartphone, digital phenotyping, social behaviour

Introduction

The appeal of a portable, ubiquitous measurement device that can monitor symptoms of mental disorders in a naturalistic environment has brought a lot of interest to the field of digital phenotyping[1,2]. This field aims to progress mental healthcare by enabling the passive monitoring of individuals with and without mental disorders, with key objectives being to quantitively measure symptom severity and to predict relapse of episodes, such as depression or psychosis. However, early excitement has been followed by a realisation that within digital phenotyping, researchers need to address important practical challenges. One of the most significant of these challenges is handling the high rates of missing data occurring in digital phenotyping datasets.

Digital phenotyping is the quantification of behavioural or physiological markers, i.e. digital phenotypes, from digital measures. The digital phenotype may reflect digital behaviours, such as phone usage, or behaviours measured from digital signals, such as step count. Whilst a range of devices can be used to carry out digital phenotyping, including smart wristbands such as Fitbit,[3] digital phenotyping using smartphones is popular due to the smartphone's already far-reaching position in modern society. Several apps exist to record digital phenotyping data on smartphones, including "Behapp",[4] "Mood mirror"[5] and "RADAR-base pRMT".[3] Early investigations in the field of digital phenotyping have found promising results, for example correlations between digital phenotyping features and depression symptom severity,[6,7] as well as age and diurnal effects.[8]

However, across studies the theme of missing data has emerged as a hurdle that is impacting many digital phenotyping research efforts,[9] and may slow down clinical implementation of digital phenotyping methods. Missingness may be due to user behaviour, such as disabling the data collection app or turning off a functionality such as GPS, or technical issues, such as changes in app permissions. As iOS or Android rules regarding what data apps are allowed to access may change, it is also possible that missingness may occur in previously successfully collected channels.

In a systematic review of digital phenotyping methods in Major Depressive Disorder (MDD),[10] it was seen that 20 out of 24 identified papers referred to missingness or acknowledged absent data. To handle missing data, samples are often excluded from analysis, however this approach may arguably only be appropriate for studies involving time-averaged comparisons. In time series analyses, dropping missing values would lead to non-consecutive datapoints being considered to be next to each other. Ten studies provided a threshold that was used to determine if a sample (e.g. participant or day) was excluded. Three studies excluded samples from their analysis if any variable was missing as the models they used did not permit missing data, however one study noted that this led to the loss of large volumes of data (89% of data).[11] Only five studies mentioned carrying out some kind of imputation on missing data, including using the mean from preceding data,[12] using a resampling method applied to GPS data,[13] using linear interpolation on Bluetooth data[14] and imputing missing records using similar complete records.[15]

Imputing missing digital phenotyping periods is valuable as this increases the number of usable samples, therefore increasing the power of the statistical analyses and broadening the number of methods available. Researchers are then not limited to only using methods that can handle missing data. However, timeseries give rise to additional challenges for imputation algorithms due to the need to model temporal dependency between timepoints. Various general methods for imputing time series exist, with popular methods including simple methods such as imputation using the mean/mode, or last-observation-carried-forward. There also exist more advanced methods that take advantage of the relationships between variables to carry out imputation. Missing values in given variables are imputed

using the available values in other variables to estimate appropriate values (for example in Multivariate Imputation by Chained Equations (MICE)[16]). Mean/mode or last-observation-carried-forward methods result in the same value being used for imputation for the given missing period. Whilst this may not cause issue if only a small number of points are missing, these methods can be problematic if the missing period is of any notable length and may warp the data distribution. Within digital phenotyping, missingness may occur across all variables at the same timepoints, meaning that multivariate imputation methods are unsuitable as there are no available points of information in any of the activity channels. This situation can occur if there is a malfunction of the data collection device.

Within smartphone-based digital phenotyping, there have been limited direct investigations of appropriate methods for imputation. One study used Gaussian Process models to impute GPS data collected using smartphones[17], and another study applied various regression models to smartphone and other sensor data to impute social media features[18]. Many of the direct investigations of imputation within digital phenotyping have occurred within wearables research, in particular for accelerometer data. Employed imputation methods include multiple imputation[19,20], neural networks,[21] and k-nearest neighbours[22,23]. Accelerometer and GPS data have high temporal resolution, contrasting with other digital phenotyping channels such as app usage, which are often more sparse. Additionally, methods that rely on other datapoints being present at a timepoint for imputation cannot be used in cases where missingness occurs across all channels at the same timepoint.

We propose the use of Poisson point process models (PPPMs) as a method to impute missing digital phenotyping data occurring in activity-based channels. A Poisson distribution can be used to model the probability of a point occurring within a given domain, based on available data. In the case of digital phenotyping, this could be the probability of an event occurring within a certain window across the time domain, for example using an app or making a phone call. In a homogeneous PPPM, the probability of an event occurring is fixed across a time series. When covariates are included in the model to allow for the probability of an event occurring to fluctuate, for example the time of day, this is referred to as a non-homogeneous PPPM.

In this study we aimed to investigate the suitability of PPPMs for imputing missing digital phenotyping data, evaluating performance in smartphone digital phenotyping datasets related to phone usage activities collected using the Behapp app [24] (Figure 1). We chose to focus on individualised models as this allowed for model parameters to be tailored towards individual patterns, therefore making comparisons within each participant, although it would also be possible to train group models.

In part one of our study, we first investigated PPPM performance against a ground truth from a selection of participants from the Smartphone based Monitoring and cognition modification Against Recurrence of Depression (SMARD) dataset[25] with high data availability and long timeseries length (months to years), providing sufficient data points for cross-validation. Covariates “time of day” and “day of the week” were initially evaluated in-sample using Likelihood Ratio Tests. We then evaluated the model generalisability out-of-sample by calculating the Poisson likelihood of the portions of data withheld in the cross-validation (which were regarded as “missing“), given the trained PPPMs, using the mean Poisson likelihood across cross-validation folds within participants to compare the models with different covariates. These pseudo-missing periods were then imputed by using the trained models to simulate point processes, where the simulations matched the pseudo-missing periods in terms of length and covariate values (Figure 1d). We then evaluated the impact the imputation had on further analysis by training Hidden Markov Models (HMMs)[26] separately on the imputed data and the ground truth data, comparing HMM parameters to evaluate whether the imputation was able to preserve properties

of the data (Figure 1e). HMMs use latent states to provide reduced dimensionality data representations, where participants transition between the different hidden states between timepoints.

In the second part of our study, we validated the use of PPPMs for imputation in a real case of missingness from the Psychiatric Ratings using Intermediate Stratified Markers (PRISM)[1] and Hersenonderzoek[27] datasets (Figure 1f). We replicated the key steps in [26] (Figure 1g), now imputing the missing data with PPPMs, to evaluate whether the imputation would impact the findings from this paper (Figure 1h).

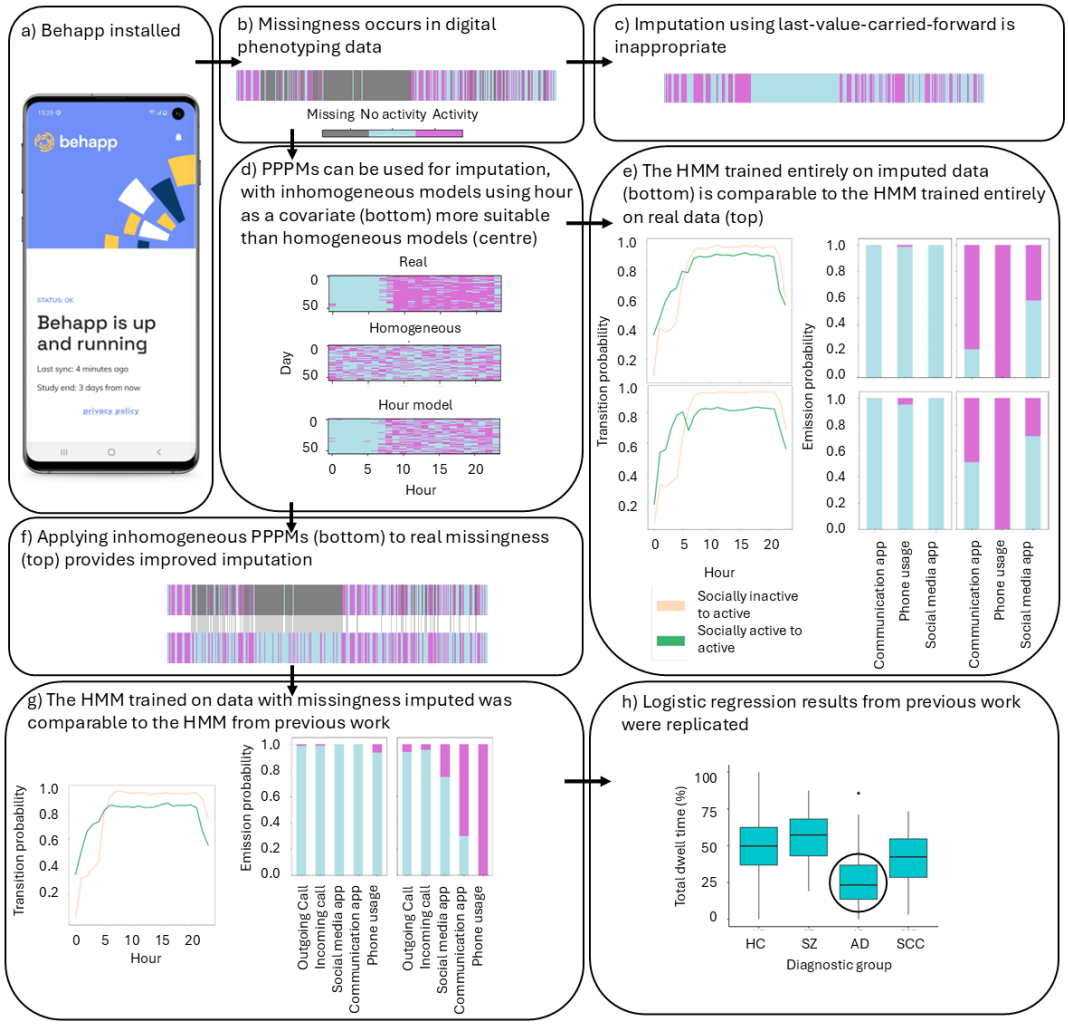

Figure 1: An overview of the imputation method used in this study. a) Behapp was installed on participants' smartphones for passive data collection, however b) this data is susceptible to missingness. c) Methods like last-value-carried-forward can be used for imputation but are not very suitable as each section of missingness is imputed using the same value. d) Poisson Point Process Models (PPPMs), in particular those using hour as a covariate, provide improved imputation. e) When compared to a Hidden Markov Model (HMM) trained entirely on real data, it can be seen that an HMM trained entirely on imputed data learnt similar model properties, including temporal properties (specifically the probability of transitioning between hidden behavioural states). f) The imputation provided from the PPPMs is more realistic than the imputation in panel c). g) In a replication analysis, a HMM trained on a dataset with missingness imputed using non-homogeneous PPPMs remained very similar to the original study, with h) previous results successfully replicated.

Results

Sample statistics

This study used data from participants from three studies: SMARD, PRISM and Hersenonderzoek. The SMARD participants were used for the first part of the study, where imputation methods were compared against the ground truth data. The PRISM and Hersenonderzoek participants were used in the second part of the study to investigate the application of the selected imputation method to real missingness, in a replication of [26].

SMARD

The SMARD study recruited participants in the Netherlands with a previous diagnosis of MDD who were classified as being in remission at the beginning of study participation[25]. From the total dataset, a subset of participants who had very good data availability (i.e. a maximum of 1% missingness) were selected. Twenty-six participants met this criterion, with a mean age of 51 (±14) years. Twenty-two of these participants were female and participants had experienced a mean number of 4.4 (±2.5) depressive episodes. Ten had completed higher vocational education, ten had completed secondary vocational education, four had completed higher scientific/academic education and two had completed intermediate secondary education. The mean time series length across participants was 213 (±214) days.

PRISM and Hersenonderzoek

The combined PRISM and Hersenonderzoek datasets contained 247 healthy controls (HCs), 18 participants with schizophrenia (SZ), 26 with Alzheimer's disease (AD) and 57 participants with subjective cognitive complaints (SCC), also referred to as memory complaints (see Table 1). Both datasets contained participants with AD and HCs, only PRISM contained participants with SZ and only Hersenonderzoek contained participants with SCC. Each study age-matched HCs to their respective diagnostic groups. For PRISM, the mean time series length across participants was 34 (±13) days. For Hersenonderzoek, the mean time series length across participants was 35 (±12) days. Further details can be found in [26].

Table 1: Demographics of the combined PRISM/Hersenonderzoek datasets.

Diagnostic group	Number	Age (Mean (SD))	Sex	Dataset	Country	Education years (Mean (SD))
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)
Subjective cognitive complaints	57	61 (7)	F=36; M=21	PRISM=0; HO=57	NL=57	5 (2)

SD: standard deviation; F: female, M: male; NL: the Netherlands, ES: Spain.

Part 1: Ground truth-based imputation investigations

To investigate the performance of the imputation methods, we required a ground truth for each timeseries. We therefore selected 26 participants from the SMARD study with very high data availability (maximum 1% of their time series missing), and used a 10-fold cross-validation approach to investigate model performance, where for each iteration 9 folds of a participant's time series were used to train each PPPM and the remaining fold was considered to be 'missing', with the actual values from each 'missing' fold used as the ground truth. PPPMs were trained for the three different available smartphone activity channels (overall phone usage, communication app usage, social media app usage). For each participant we trained homogeneous PPPMs and non-homogeneous PPPMs where 'hour of the day' and 'day of the week' were separately used as covariates. Due to the cyclical nature of the covariates, the covariate values, which ranged from 0 (midnight) to 23 (11pm) for hour and 0-6 for day, were encoded to remove the ordinal implications that 11pm was the furthest value away from midnight (and likewise for the other values). We investigated two different methods: one-hot encoding and sine-cosine transformations. Further explanation of these two encoding methods are provided in the Methods. For participants where the 'day of the week' covariate was significant based on likelihood ratio tests (LRTs), additional PPPMs that included both hour and day were also trained.

In-sample likelihood ratio tests

To investigate which covariates improved the fit of the PPPMs, LRTs were used to compare the non-homogenous PPPMs to the baseline homogeneous PPPM. A *p*-value threshold of .05 was used to determine whether each respective covariate was a significant contributor to the model. For each activity channel, LRTs were run for each trained model (i.e. each participant/fold combination), with comparisons made separately between the baseline homogeneous model and models that each included one of the following covariates: one-hot encoded hour, sine-cosine transformed hour, one-hot encoded day and sine-cosine transformed day. The percentage of models of all cross-validation runs for which each covariate was a significant contributor are provided in Table 2. For almost all models the hour was significant, with the lowest percentage of models for which this was significant across the channels being 98%. With respect to the day covariate, this was less commonly a significant covariate, with the maximum percentage of models for which this was significant at 30% (for the communication app channel using one-hot encoding). The day covariate appeared least useful for the overall phone usage channel. To ensure that the day was including additional information relative to the hour, for the participants where the day covariate was significant for at least one of their folds (and considering each encoding method), additional models were trained that included both day and hour as covariates. LRTs were then run to compare the models including both day and hour as covariates to models solely including the hour (
Table 3). Day generally remained a significant contributor to the PPPMs in within-sample comparisons.

Table 2: Percentage of Poisson Point Process Models for which each covariate is significant according to likelihood ratio tests (total number of models per row=260). Note that 14/21 of the models that failed to converge were from 1 participant. n=number.

Channel	Transform	Significant day models (% (n))	Significant hour models (% (n))	Day models that failed to converge (% (n))	Hour models that failed to converge (% (n))
Phone usage	One-hot	8 (21)	99 (258)	1 (3)	1 (2)

Phone usage	Sine-cosine	13 (34)	100 (260)	0	0
Communication app	One-hot	30 (79)	99 (257)	1 (3)	1 (3)
Communication app	Sine-cosine	23 (61)	100 (260)	0	0
Social media app	One-hot	27 (71)	98 (254)	2 (4)	2 (5)
Social media app	Sine-cosine	16 (42)	98 (255)	0.4 (1)	0

Table 3: Likelihood ratio test results comparing models with both day and hour to models solely with hour (for participants where the day covariate was significant).

Channel	Transform	Significant combined day and hour models/significant hour models (%)
Phone usage	One-hot	9/21 (43)
Phone usage	Sine-cosine	33/34 (97)
Communication app	One-hot	63/79 (80)
Communication app	Sine-cosine	60/61 (98)
Social media app	One-hot	60/71 (85)
Social media app	Sine-cosine	42/42 (100)

Out-of-sample Poisson likelihood

In order to assess generalisability, in the SMARD dataset the out-of-sample Poisson log-likelihood was then calculated for each participant across each of their left out cross-validation folds, using the corresponding PPPM trained on the other 9 folds. This likelihood was then averaged to provide a mean out-of-sample Poisson loglikelihood per participant for each PPPM type. The log-likelihoods for each of the channels can be seen in Table 4, Table 5 and Table 6. It can be seen that the Poisson log-likelihoods for the non-homogeneous models that include hour as a covariate are almost always higher than the likelihood from homogeneous Poisson models, though for some participants the likelihoods across models are similar. In addition, the models that included hour encoded using one-hot encoding were an improvement upon the models where hour is encoded using sine-cosine transformation. As expected given the in-sample LRTs, the inclusion of day as the only covariate did not improve the out of sample log-likelihood. For the participants with models including both day and hour, the log-likelihood was rarely an improvement upon the models including only hour, with the models including both day and hour providing the highest mean Poisson likelihood for 0/7 possible participants for phone usage, 5/13 possible participants for communication app channel (with 2 of these negligibly higher), and 2/14 possible participants for the social media app channel (with 1 of these negligibly higher). We therefore decided to use PPPMs that included one-hot encoded hour as our general imputation approach for the remainder of our analyses (Figure 1d).

Table 4: Mean out-of-sample log-likelihoods for the overall phone usage channel. The highest likelihoods are highlighted.

Participant	Homogeneous	Hour (one-hot)	Hour (sine-cosine)	Day (one-hot)	Day (sine-cosine)	Hour+Day (one-hot)	Hour+Day (sine-cosine)
1	-776	-683	-733	-776	-776		
2	-1126	-898	-1002	-1125	-1125		-1002
3	-1334	-1298	-1262	-1334	-1333	-2495	-1262
4	-117	-111	-114	-117	-117		
5	-467	-383	-439	-467	-467		
6	-1419	-1329	-1351	-1419	-1419		
7	-1403	-1250	-1321	-1403	-1403		
8	-956	-735	-825	-957	-956	-738	
9	-308	-266	-287	-308	-308		
10	-126	-116	-120	-126	-126		
11	-143	-134	-135	-143	-143		
12	-62	-56	-58	-62	-62		-58
13	-136	-117	-124	-136	-136		
14	-260	-215	-240	-260	-260		-240
15	-551	-464	-500	-550	-550		
16	-16	-15	-13	-15	-16	-44	
17	-129	-111	-119	-130	-129		
18	-62	-56	-59	-62	-63		
19	-129	-116	-119	-129	-129		
20	-281	-231	-251	-281	-281		
21	-50	-48	-47	-51	-50		
22	-300	-262	-282	-300	-300		
23	-280	-227	-246	-280	-280		
24	-110	-97	-103	-110	-110		
25	-365	-307	-341	-365	-365		
26	-1214	-1073	-1144	-1214	-1214	-1096	-1143

Table 5: Mean out-of-sample log-likelihoods for the communication app channel. The highest likelihoods are highlighted.

Participant	Homogeneous	Hour (one-hot)	Hour (sine-cosine)	Day (one-hot)	Day (sine-cosine)	Hour+Day (one-hot)	Hour+Day (sine-cosine)
1	-694	-616	-655	-694	-694		
2	-938	-779	-840	-937	-937	-770	-839
3	-1168	-1136	-1110	-1168	-1167	-972	-1109
4	-116	-109	-112	-116	-116		
5	-417	-347	-393	-417	-417		-393
6	-1268	-1163	-1193	-1268	-1268		
7	-1294	-1136	-1200	-1293	-1294	-1141	

8	-857	-676	-748	-857	-857	-675	-748
9	-301	-261	-281	-301	-301		
10	-112	-104	-107	-112	-112		-107
11	-126	-118	-118	-126	-126		
12	-60	-54	-57	-60	-60	-54	-57
13	-127	-111	-116	-127	-127		
14	-239	-201	-221	-238	-238	-200	-220
15	-514	-438	-471	-514	-514	-438	-470
16	-14	-14	-13	-13	-14	-46	
17	-112	-91	-101	-112	-112		
18	-57	-50	-53	-57	-57		
19	-111	-100	-102	-111	-111		
20	-265	-224	-237	-265	-265		
21	-43	-38	-39	-43	-43		
22	-233	-201	-216	-233	-233	-201	-216
23	-270	-213	-230	-270	-270		
24	-100	-92	-94	-101	-100		
25	-319	-265	-297	-319	-319	-267	
26	-1188	-1051	-1121	-1187	-1187	-986	-1121

Table 6: Mean out-of-sample log-likelihoods for the social media app channel. The highest likelihoods are highlighted.

Participant	Homogeneous	Hour (one-hot)	Hour (sine-cosine)	Day (one-hot)	Day (sine-cosine)	Hour+Day (one-hot)	Hour+Day (sine-cosine)
1	-381	-360	-370	-381	-381	-360	-370
2	-706	-592	-638	-707	-706		
3	-846	-901	-835	-847	-846	-815	-835
4	-95	-90	-94	-95	-95		
5	-202	-179	-200	-203	-202	-180	-200
6	-781	-722	-740	-782	-781	-723	-740
7	-728	-675	-701	-728	-727		
8	-908	-712	-800	-908	-908	-723	
9	-168	-157	-162	-168	-168		
10	-89	-84	-88	-90	-90		
11	-92	-90	-88	-92	-92		
12	-37	-35	-36	-38	-37	-36	
13	-78	-73	-71	-78	-78		
14	-139	-120	-129	-139	-139	-120	-129
15	-234	-210	-216	-233	-234	-209	-216
16	-7	-10	-7	-6	-7	-14	
17	-109	-91	-100	-109	-109		
18	-22	-27	-21	-23	-22		-22

19	-100	-95	-97	-100	-100	-95	-97
20	-231	-199	-213	-231	-231		
21	-44	-42	-42	-44	-44		
22	-277	-247	-263	-276	-276	-247	-263
23	-86	-76	-77	-87	-87	-76	-78
24	-72	-68	-70	-73	-73		
25	-132	-123	-129	-132	-132		
26	-1010	-924	-983	-1010	-1011	-924	-984

Visualisation examples

Next, we provide a set of visualisations in order to visualise the effect of covariates on the imputation quality. For each of the participants we simulated the 'missing' periods, i.e. each withheld fold, using the PPPM including one-hot encoded hour that was trained on the remaining folds. The simulated time series matched each withheld fold in terms of length and covariate values. We present the simulated visualisations alongside the real time series visualisations for the corresponding folds for the phone usage channel in Figure 2. Figure 2a and 2b display participants who had very clear diurnal rhythm with very low activity levels during nighttime, and 2c shows a participant with more elevated activity levels during the nighttime. It can be seen that the PPPMs including hour as a covariate provided improved imputation when compared to the homogeneous PPPMs, capturing the personal diurnal behaviours of each participant. The corresponding figures for these participants for communication app usage and social media app usage are provided in the Supplementary Materials (Figures S1-S2).

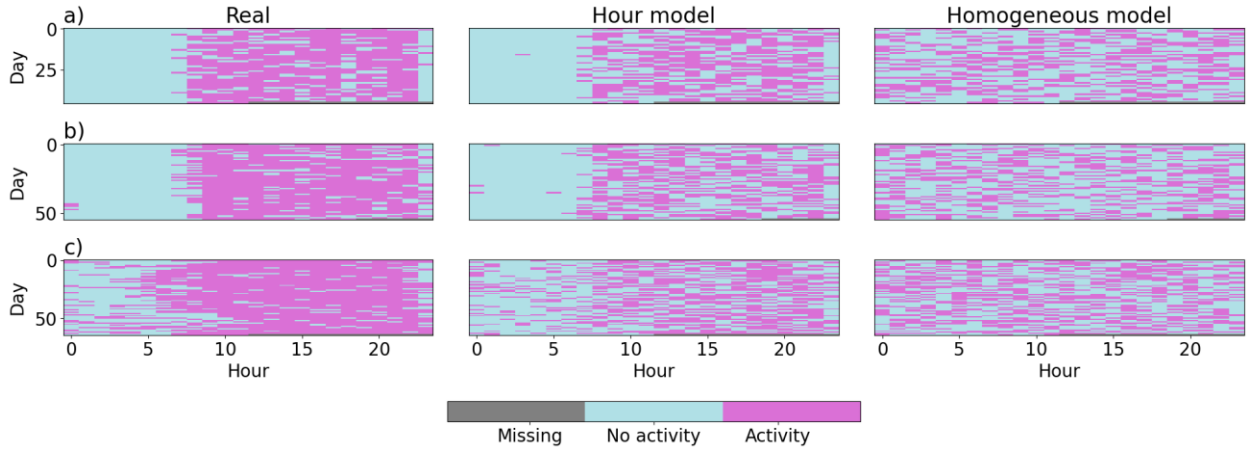

Figure 2: Real versus imputed time series for the phone usage channel for three participants (a), b) and c)). Each row corresponds to a single fold from each participant from their cross-validation. The left column shows the fold's original time series, the centre column shows the simulation provided by the PPPM using one-hot encoded hour, and the right column shows the corresponding homogeneous PPPM. Note that the fold's original time series was not used in model training. It can be seen that for each participant the hour model captured individual behaviour more accurately than the homogeneous model. The hour models for participants a) and b) show notably lower activity during the nighttime, as expected given their real time series. The hour model for participant c) shows higher levels of nighttime activity, reflecting this participant's real behavioural pattern.

Hidden Markov Model comparisons between real and imputed datasets

The simulated SMARD data was then concatenated and used to train HMMs which were compared to an HMM trained entirely on the real SMARD data (Figure 1e). As activity points can only be present in the communication app and social media app channels when there is activity present in the overall phone

usage channel, for the simulated data we investigated two HMM types. The first HMM was trained on the original simulations for each activity channel, and the second HMM had the imputed communication app and social media app channels constrained based on the imputed phone usage channel, where the imputed channels were set to contain no activity at the timepoints where imputed phone usage showed no activity. The latter is the more plausible behavioural pattern given the dependence of the other channels on overall phone usage, although this of course reduces the number of activity points.

It can be seen from the emission probabilities in Figure 3a that the HMM trained on the real data has one state with very low probabilities of activity across channels, and another state with a very high probability of activity in the phone usage channel, high probability of communication app usage and then lower probability of social media app usage. The HMM trained on the imputed time series (Figure 3b), without constraining based on phone usage activity, also shows low activity and high activity states, however the high activity state here gives lower probabilities than the HMM trained on the real data, with phone usage still the most probable channel and social media app the least. Once the imputed time series is constrained so that no activity is present in the channels if there is no phone usage activity (Figure 3c), the probability of phone usage is more comparable to the HMM trained on the real data.

The transition probabilities for each of the HMMs are shown in Figure 4. It can be seen that for all HMMs, the probability of transitioning into the high activity state, from either state, rapidly increases during the morning, is high throughout the day, and then decreases in the evening.

Regarding the starting probabilities, for both HMMs trained on imputed data, there was a 1.0 probability of starting the time series in the inactive state. This contrasts with the HMM trained on the real data, where the probability of starting the time series in the inactive state was 0.44.

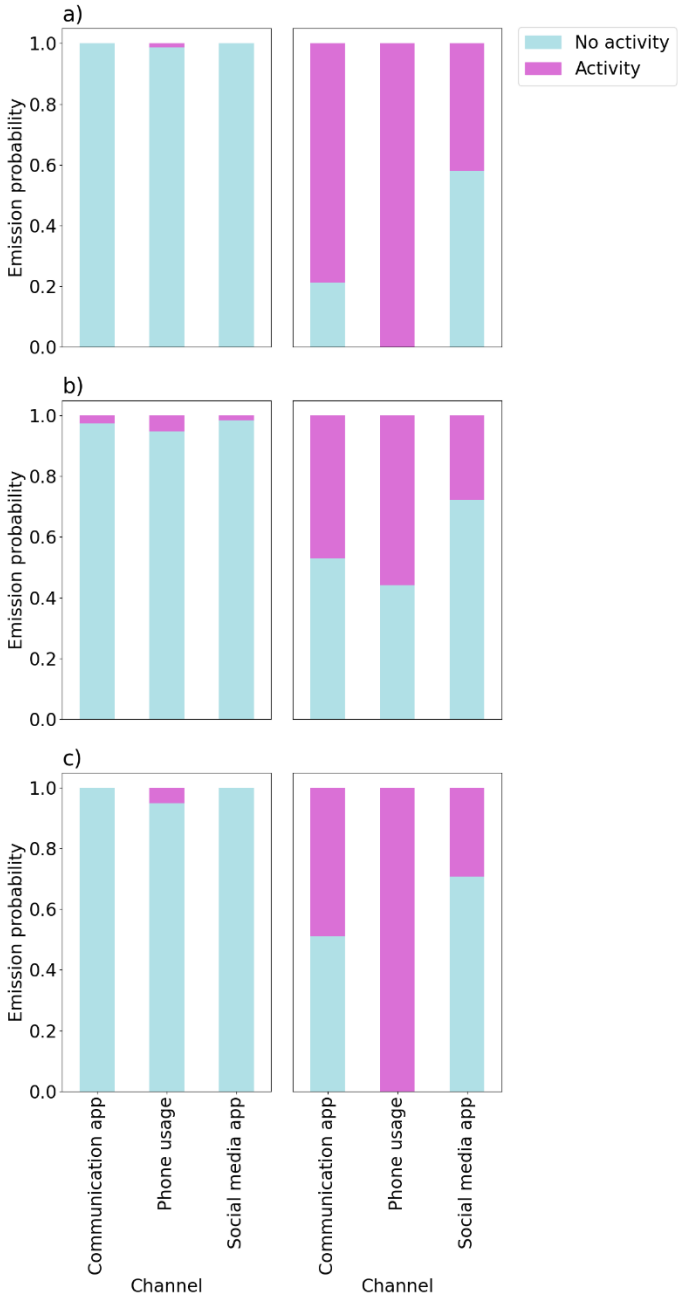

Figure 3: Emission probabilities from the HMMs trained entirely on a) real data; b) imputed data, where channel values were unconstrained by phone usage; c) entirely on imputed data, where channel values were constrained by phone usage. It can be seen that each HMM has a low activity state (lefthand side) and a state with higher probabilities of activity (righthand side). The high activity states are referred to as 'socially active' and the low activity states are referred to as 'socially inactive', for consistency with the second part of this study.

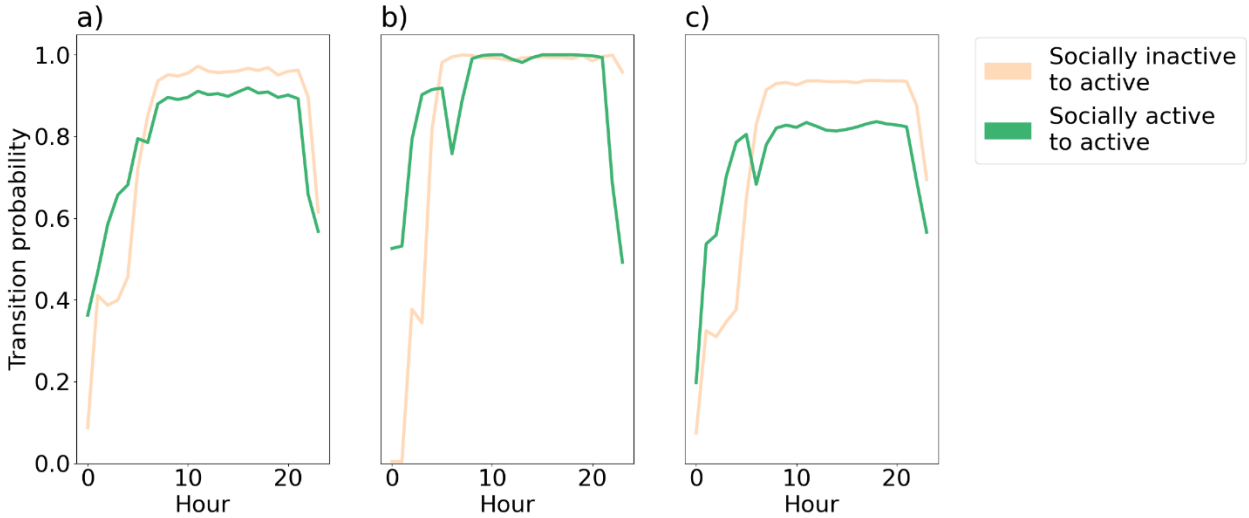

Figure 4: Transition probabilities between hidden states from the HMM trained entirely on a) real data; b) imputed data, where channel values were unconstrained by phone usage; c) imputed data, where channel values were constrained by phone usage. The high activity states are referred to as 'socially active' and the low activity states are referred to as 'socially inactive'.

Part 2: Imputation application to real missingness

Training a Hidden Markov Model on a dataset where missingness was imputed using PPPMs

Following the ground truth-based investigations, we sought to apply the imputation method to a case of naturalistic missingness occurring in the PRISM and Hersenonderzoek datasets (Figure 1f). We repeated the key parts of the analysis in the original paper, where a HMM was trained on the available Behapp channels: phone usage, communication app, social media app, incoming and outgoing calls. A measure derived from the HMM, referred to as the total dwell time, was then compared between the different diagnostic groups [26]. Our key finding from [26] was that participants with AD, as well as participants with memory complaints, had lower total dwell time than HCs (i.e. they spent less time in the socially active hidden state), with the HMM from the original study referred to as 'HMM-original'. We sought to replicate the hidden states identified in [26], training the HMM on an updated time series where missingness was imputed using PPPMs, and to evaluate whether we would still identify these significant group differences.

Before training the HMM, missing periods were imputed using non-homogeneous PPPMs trained for each participant (Figure 1f). Given the results of the out-of-sample likelihood investigations, we chose to use PPPMs that included one-hot encoded hour of the day as the sole covariate in the model. We trained a PPPM for each participant on their entire available data (excluding their missing periods) and used this model to impute their missing periods by taking the corresponding values from the simulated time series provided by their PPPM. The percentage of missing datapoints in the combined PRISM and Hersenonderzoek dataset decreased from 30% to 4% after imputation, with the 4% missingness from participants where imputation was not reasonable (see Methods). We constrained the imputations in the other channels based on imputed phone usage, so that values were only allowed in other channels at timepoints where phone usage was predicted. Example real time series and their corresponding simulated time series used to impute missing periods are shown in Figure 5. We then

trained a 2-state HMM on the whole PRISM and Hersenonderzoek combined dataset using their updated time series (referred to as 'HMM-imputed') (Figure 1g).

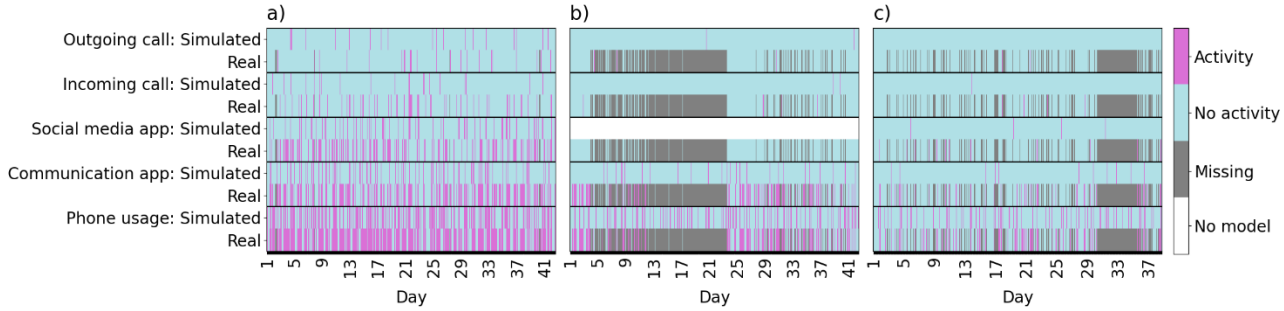

Figure 5: Example real time series from three participants (a), b) and c)) with their corresponding simulated time series provided by their individual Poisson Point Process Models, which was used to impute the missing grey periods. a) A participant with very low missingness. The simulated time series show similar patterns, most obvious in the phone usage channel, but underestimate the number of activity points; b) a participant with notable missingness; c) a participant with moderate missingness.

Figure 6 shows that the emission probabilities from HMM-imputed are extremely similar to the emission probabilities from HMM-original, meaning that the interpretation of each hidden state as 'socially active' or 'socially inactive' (as in the original study) is also reasonable for the HMM trained on data where missingness was imputed. The transition probabilities also show the same pattern of an increased probability of transitioning into the socially active state during the daytime from either hidden state (see Figure 7).

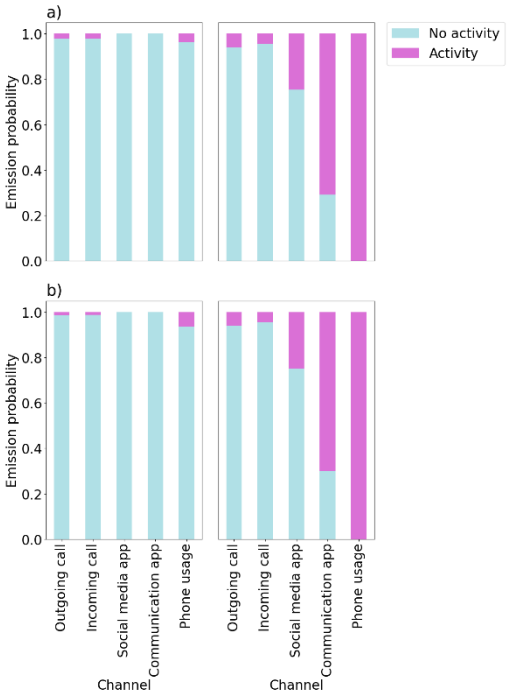

Figure 6: Emission probabilities for a) the two state HMM from the original paper, trained using all participants and b) the HMM trained in the same way except with missing data imputed (and constrained based on phone usage). The states on the lefthand side of a) and b) show a very low probability of activity, whereas the righthand side shows higher probabilities of phone and app usage. Differences between a) and b) are very small.

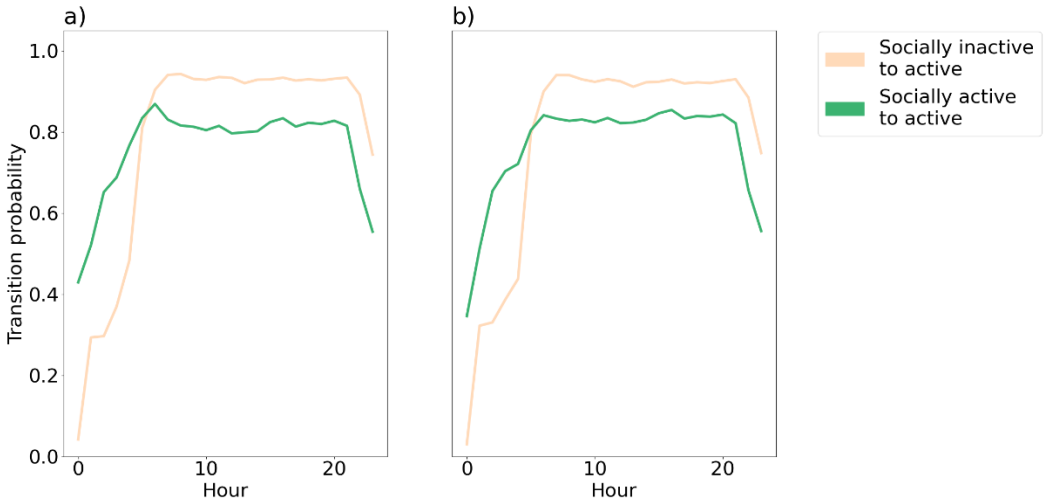

Figure 7: Transition probabilities for a) the two-state HMM from the original paper, trained using all participants and b) the HMM trained in the same way except with missing data imputed (and constrained based on phone usage). For both a) and b), the probability of transitioning into the socially active state from either the socially active or inactive state is higher during the daytime than the nighttime. Differences between a) and b) are small. 0: midnight.

Differences between diagnostic groups

Finally, as in the prior study we investigated the difference in the time spent in the socially active state between the different diagnostic groups and HCs using multinomial logistic regression. The multinomial logistic regression model using data where missingness was imputed (HMM-imputed) also replicated the results from the original model estimated on real data (HMM-original) (Figure 1h), with the time spent socially active significantly lower in the AD and SCC groups relative to HCs (see Figure 8 and Table 7 for results from the total dwell time predictor and Table S1 in the Supplementary Materials for the results from the age predictor).

Table 7: Multinomial logistic regression results from the total dwell time predictor, using the HMM that had imputation constrained based on phone usage. SZ: schizophrenia, AD: Alzheimer's disease, SCC: subjective cognitive complaints.

Group	Coefficient	Standard error	Odds ratio	*z* value	*P* value	FDR corrected *P* value
SZ	-0.0173	0.0158	0.9829	-1.0962	0.27	0.82
AD	-0.0530	0.0128	0.9484	-4.1369	<.001	<.001
SCC	-0.0232	0.0080	0.9771	-2.9196	0.004	0.01

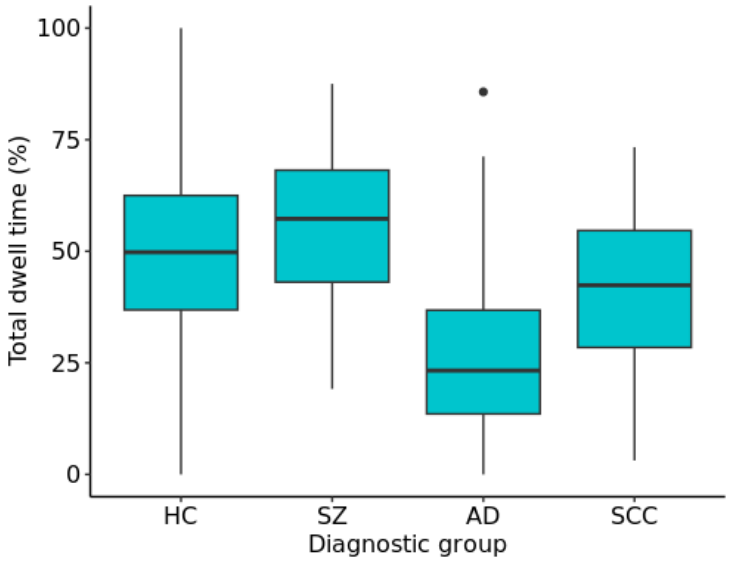

Figure 8: A box plot of the total dwell times per participant for the different diagnostic groups, using the HMM that had imputation constrained based on phone usage. There is a significant difference between the HC and AD groups, and the HC and SCC groups. HC: healthy control, SZ: schizophrenia, AD: Alzheimer's disease, SCC: Subjective cognitive complaints.

Discussion

Principal findings

In this study we demonstrate the utility of Poisson Point Process Models (PPPMs) as a method for imputing missing smartphone time series data, where smartphone activity is conceptualised as a point that occurs with a probability learnt by the model. We investigated various PPPMs and selected the PPPM using one-hot encoded hour as the most suitable, using this imputation method in a successful replication of a prior digital phenotyping study using Hidden Markov Models (HMMs).

We carried out ground-truth based investigations using participants from the SMARD study[25], where we masked periods as missing and imputed these periods using models trained on the remaining periods. We investigated homogeneous PPPMs as well as models that included the hour of the day and day of the week as covariates, training individual models so that the probability that was learnt was personalised to each individual. When considering the out-of-sample likelihood we identified that the non-homogeneous models that included one-hot encoded hour of the day as a covariate generally provided the best models, suggesting that accounting for diurnal variation is important for imputation. When sine-cosine transformed hour was used, PPPMs appeared to overestimate nighttime activity and underestimate daytime activity. However, during model training we found that using sine-cosine transformation as the encoding method provided more consistent convergence, with model training using one-hot encoding requiring more tailoring for participants where convergence failed. The non-homogeneous PPPMs using one-hot encoded hour were used to simulate the time series, and then HMMs were trained on the 'imputed' and the real time series. The model parameters from both of these HMMs were comparable, suggesting that the imputed time series preserved the properties of the original data.

We then investigated the influence of applying imputation using PPPMs with one-hot encoded hour of the day to a prior digital phenotyping study that used HMMs to investigate group differences between various clinical populations and HCs. When imputation was used to fill the missing periods, the

learnt HMM had comparable model parameters to the original HMM. We subsequently demonstrated significantly lower time spent socially active for participants with AD and memory complaints compared to HCs, consistent with the previous study that included no imputation.

Future directions

In the current study we investigated various covariate and encoding options and then selected the generally best performing model to use for imputation. However, given that the PPPMs are already trained per person, it would be possible to carry out individualised covariate selection, so that different participants have different covariates used in their imputation models. For real world cases this selection would have to be carried out in-sample, and so should be done conservatively. For example, in the in-sample comparisons for SMARD participants it was seen that the day of the week may be a useful covariate to include for several participants. However, when then making out-of-sample comparisons, including the day was infrequently an improvement upon using solely the hour, and only for communication app or social media app channels. Additional covariates that may affect phone usage behaviour could be explored, such as weather data, or location data (e.g. periods spent at home) if available. The relevant covariates used may depend on the channel being imputed, for example the weather may impact someone's usage of maps-related apps.

The non-homogeneous PPPM could be applied to further data types. For example, to typing data or to ecological momentary assessment data. The PPPM method could also be used to infer whether a person has diurnal and/or weekly behavioural patterns as a question of interest in its own right, or to determine whether certain covariates have a meaningful influence on digital phenotyping signals. The suitability of using PPPMs to predict digital phenotyping points that have an associated continuous value could also be investigated, e.g. percentage of an hour containing activity as opposed to activity versus no activity. In this case marked PPPMs could be used to predict the associated percentages.

In our ground truth HMM investigations we compared the real data to an extreme missingness case where all data was imputed, as each participant's withheld folds were concatenated to provide an entirely imputed time series. This is of course not feasible, and future work could seek to investigate the limits to how much data can be reliably imputed using PPPMs in a realistic setting.

Limitations

In our study we treated a homogeneous PPPM as our baseline, however to our knowledge, this in itself is a novel imputation method for digital phenotyping data. We made this decision as several other methods had obvious limitations (as using the mode or last-observation-carried-forward would mean only one value was used to impute missing periods, potentially altering the data distribution) or could not be applied to our data (as is the case for multivariate imputation methods). It would have been preferable, however, to have been able to use a more established baseline.

Whilst our non-homogeneous PPPM imputation method can learn individualised behavioural patterns, our approach of using PPPMs applied to each smartphone activity channel does not account for covariation between the different channels. In our case, activity across all channels is constrained by the phone usage channel, and it could also be possible for behaviours related to one channel to affect other channels. Future research could seek to investigate multivariate imputation using hierarchical or multivariate PPPMs.

Whilst our datasets have been prone to chunks of missingness, we do see that in some participants missingness occurs periodically (see Figure 5c). In these cases, the parameters learnt by the model to reflect the covariate values missing at these timepoints are unlikely to be a good

representation of the participant's true behaviours due to the reduced sample size available for model training for these covariate values. In these cases it may be useful to investigate whether including parameter values from other participants with similar time series may improve imputation. It is also possible that recording additional user behaviours (e.g. phone switch offs) may explain this missingness and be useful information in and of itself.

As our datasets include several clinical populations (MDD, AD, memory complaints, SZ), it is possible that their time series include behavioural changes that can be attributed to their symptom progression and may affect imputation. For our initial investigations we wanted to train with as much data as possible for each participant, however for further scientific investigations it could be evaluated whether to use all data for imputation or whether to select periods with certain criteria, e.g. low symptom severity, therefore excluding periods where behaviour may be impacted by high symptom severity.

Conclusion

Digital phenotyping datasets are prone to missingness, being one of the key challenges for the field to address. We proposed the use of non-homogeneous PPPMs as a tool for imputing missing activity in smartphone time series. We found that the non-homogeneous PPPMs using hour of the day as a covariate encoded using one-hot encoding provided the highest out-of-sample Poisson likelihood and therefore selected this model type as the model to be used for imputation. When comparing HMMs trained entirely on imputed data versus trained entirely on ground truth data, the HMM parameters remained similar. Additionally, we replicated the key results from a previous HMM study with the missingness imputed using the selected PPPM type. Using non-homogeneous PPPMs for digital phenotyping imputation is a promising avenue for improving the analysis and utility of digital phenotyping data.

Methods

Datasets

SMARD dataset

The SMARD study[25] aims to investigate recurrence of depressive episodes in a population with a classification of recurrent MDD (according to the Structured Clinical Interview for Diagnostic and Statistical Manual of Mental Disorders, fourth edition (DSM-IV) (SCID)).[28] Participants were required to be in stable remission at the beginning of the study (Hamilton Depression Rating Scale score ≤10 for ≥8 weeks),[29] to have a vulnerability for recurrence (≥3 prior depressive episodes), be between 18 to 65 years of age and to be a regular smartphone user. Exclusion criteria included a diagnosis of bipolar, primary psychotic or borderline personality disorder, or primary diagnosis of substance use or anxiety disorder with a secondary MDD diagnosis. Participants were excluded if they had had electroconvulsive therapy in the two months before inclusion, an average daily alcohol intake exceeding three units, daily use of benzodiazepines, or if their smartphone was incompatible with the Behapp app. There were no in-/exclusion criteria for (stable) antidepressant usage.

Participants were invited to download Behapp on their smartphones, which passively collected smartphone measures throughout the 1.5 year study-course. The data collection for the participants used in this analysis took place between August 2019 – March 2023. Participants filled out baseline questionnaires and completed one week of experience sampling prior to Behapp data collection.

Additionally, participants were assessed every three months for depression recurrence using the SCID-I and questionnaires. Due to the long study duration, each participant had a long time series which was suitable for training individual models in this analysis.

PRISM and Hersenonderzoek datasets

The datasets used in the replication analysis were from the PRISM[30] and Hersenonderzoek[27] studies, which collected Behapp data for 42 days per participant. PRISM data collection took place between August 2017 – May 2019 and Hersenonderzoek data collection took place between March 2018 – January 2020. A more detailed explanation of these studies and their inclusion criteria is provided in the original analysis [26]. Both of these datasets were combined due to their overlap in study populations (AD and HCs).

PRISM

The PRISM study investigates disorder-related social withdrawal in SZ and probable AD,[31,32] with HCs age- and sex-matched to both clinical populations. Recruitment took place across sites in the Netherlands (University Medical Center Utrecht, Leiden University Medical Center and Amsterdam UMC, location VUmc) and Spain (Hospital General Universitario Gregorio Marañón and Hospital Universitario de La Princesa, in Madrid). All participants self-reported social functioning according to the Social Functioning Scale (SFS)[33] and loneliness according to the De Jong Gierveld Loneliness and Affiliation Scale.[34]

Participants with SZ had a DSM-IV diagnosis of SZ confirmed by the Mini-International Neuropsychiatric Interview (MINI), a minimum of one psychotic episode in their history throughout a maximum disease duration of 10 years since initial diagnosis, a positive symptom factor score of <22 from the positive symptom factor of the 7-item Positive and Negative Syndrome Scale (PANSS)[35], stable antipsychotic medication dosage for at least 8 weeks, and were between 18-45 years of age (inclusive).

Participants with 'probable AD' were classified as such based on the National Institute on Aging and the Alzheimer's Association (NIAAA) criteria, and were required to have a Mini-Mental State Examination (MMSE)[36] score of 20-26 and be between 50-80 years of age.

HCs were matched to the SZ and probable AD age ranges, and their MMSE score compared to the average for their age and years of education so that only HCs of approximately average cognitive functioning were included. Participants meeting the criteria for an Axis-I psychiatric disorder (assessed by the MINI) or a neurological disease associated with cognitive impairment were excluded.

Hersenonderzoek

Participants self-reported the presence of probable AD, or an absence of neurological or psychiatric diseases with memory complaints ('subjective cognitive complaints' (SCC)) or without memory complaints (HCs) via The Dutch Brain Research Registry's 'Hersenonderzoek.nl' platform.[27] This platform was also used to report other health-related and demographics information. Participants were recruited across the Netherlands and had a minimum age of 45 years.

Ethical considerations

SMARD was approved by the Radboudumc ethics review board (2016-3009; NL60033.091.16). PRISM was approved by the Ethical Review Board University Medical Centre of Utrecht (17-021/D) for the involved Dutch research centres, and by the Comité Ético de Investigación Clínica Hospital General Universitario Gregorio Marañón (59359) for the Spanish research centres. Hersenonderzoek was

approved by the Ethical Review Board VU University Medical Centre (2017.254). All SMARD, PRISM and Hersenonderzoek participants provided informed consent before participating in their respective study. In all studies participants were offered reimbursement for their travel expenses. PRISM participants were additionally given compensation for their time, and SMARD participants were compensated €40 upon completion of their 1.5 year follow-up.

Behapp acquisition

At the beginning of study participation 'Behapp' was installed on each participant's smartphone.[24] As earlier versions of Behapp were only available on Android smartphones, two PRISM participants who did not have their own Android smartphones were provided with one throughout the study. Hersenonderzoek-provided Android phones were not offered due to its study design, nor were SMARD-provided phones offered.

The start and end times for each activity carried out by a participant, such as use of an app, were recorded. Each app opened by a participant was classified according to its label in the Google Play Store, for example 'social media' or 'communication' app. Data related to calls was collected for PRISM and Hersenonderzoek participants, but not for SMARD participants due to an initial technical malfunction in Behapp. No content of messages or calls was collected, and collected data was de-identified, in compliance with the GDPR.[37] Whilst Behapp also collects GPS data, here we focused on the phone usage-related channels as these event-based channels are more suitable to be modelled as points with a PPPM.

Preprocessing

Smartphone channels

For SMARD, PRISM and Hersenonderzoek, smartphone behavioural activities reflecting social media app usage, communication app usage and overall phone usage were recorded. Outgoing and incoming calls were also recorded for PRISM and Hersenonderzoek. These activities were split into hourly bins giving the percentage of each hour during which each respective activity was carried out. Nevertheless, the distributions for these activities are highly zero-inflated (see [26] for more details), and so instead of using the raw percentages per hour we considered the binary values of whether an activity was carried out in each hour or not (0% activity; >0% activity). We feel that this low threshold for activity is justified given that many meaningful activities only require a short duration to carry out, such as reading or replying to a message. The first day of each participant's time series was removed in case of any behavioural changes occurring as a consequence of study onboarding, meaning each participant's time series started at midnight.

Measures of missingness

In order to differentiate between zero values and missing values, we developed a measure to identify missingness. This measure used the expected sampling frequency (>1/hour) of sensors including GPS and WiFi, which do not depend on active phone usage, to determine if data were missing; if these sensors did not provide any datapoints for an hour then this hour was considered to be missing. Consequently, for all activity channels included in this paper, any missing points occurred for the same time points for all channels. Whilst some missingness appeared periodically, these data were prone to seemingly random chunks of missingness (Figure 5c).

Poisson point process models

A homogeneous PPPM models the probability (or 'intensity') of a point occurring, with this probability being constant. In a non-homogeneous PPPM, the learnt probability can vary according to the values of covariates. During model training, a parameter for each covariate as well as the intercept is learnt through maximum likelihood estimation. We used the 'NHPoisson' package (version 3.3),[38] available in R, for PPPM training, covariate selection and time series simulation. Training was carried out using the 'fitPP' function. The Poisson loglikelihood of a series of points given the learnt model is given by the equation:

Equation 1

$$LL(\beta; (t_i)_{i=1}^{n}) = -\sum_{t=1}^{T} \lambda(t;\beta) + \sum_{i=1}^{n} \log \lambda(t_i;\beta)$$

Equation 2

$$\lambda(t;\beta) = \exp\left(\mathbf{X}^{\mathrm{T}}(t)\beta\right)$$

Where β is the vector of model parameters from the trained model, t is each individual timepoint, n is the total number of points with activity within the given time series, $\mathbf{X}^{\mathrm{T}}(t)$ is the vector of time-varying covariates at time t and T is the length of the time series. The first part of the equation therefore is calculated across the whole time series, and the second part of the equation is calculated for the timepoints with activity. To fit the model, the negative of the Poisson loglikelihood was minimised using NHPoisson's default function, 'nlminb'.

Ground truth-based imputation investigations

10-fold cross-validation was used to evaluate the suitability of each of the models, with the time series for each participant split into ten equal-length consecutive segments. Nine of these segments were then used to train the model for each participant, with the remaining segment then treated as the 'missing' portion of the data. Each of the trained models, described further below, were then used to simulate time series equal in length and covariate values to the 'missing' fold, thereby providing imputed values for this portion. The imputed time series could then be compared to the real values for this fold.

Baseline method

Homogeneous PPPMs were treated as the baseline method, where the probability of an event occurring is constant throughout the time series (i.e. no covariates are used). This type of model was chosen as the baseline as many other common imputation models would be nonsensical (Figure 1c).

Implementation of non-homogeneous Poisson point process models

Both of the possible covariates of interest, 'hour of the day' and 'day of the week', are cyclical, and therefore must be encoded in a way that does not imply a misleading order to the variable values. For example, encoding hours as ranging from 0-23 would imply that 11pm (23) and midnight (0) are far away from each other, instead of just one hour apart. We considered two methods for encoding these covariates: one-hot encoding and sine-cosine transformation. One-hot encoding involves converting each covariate value into an array of length n (number of possible covariate values) minus 1, i.e. length 6 for day of the week and 23 for hour of the day. An array of zeros is used where for each covariate value

one of the positions is set to 1, with the exception of one value whose corresponding array is entirely zeros. Sine-cosine transformation instead requires two dimensions to encode each value, v, where the value is transformed using sine and cosine functions:

Equation 3

$$x = \sin\left(\frac{v}{p} \times 2\pi\right)$$

Equation 4

$$y = \cos\left(\frac{v}{p} \times 2\pi\right)$$

Where p is the relevant period (e.g. $p = 24$ for the hour covariate).

Covariate selection

Likelihood ratio tests (LRTs) were used to select which covariates provide useful additional information for the models. LRTs compare the in-sample-likelihood between two nested models. We treated the homogeneous Poisson model as the more general model and a model with covariates as the constrained model. We ran LRTs comparing the homogeneous Poisson models to models including the day as a covariate and to models including the hour as a covariate, investigating each encoding method (sine-cosine transformation and one-hot encoding). The *p*-value from the LRT was used to assess whether the model with covariates provided a significantly better fit, with a threshold of $p<.05$ used to evaluate statistical significance.

Additionally, for participants where both the hour and day were individually significant (for each covariate encoding method), additional models trained using both the hour and the day were considered, with LRTs carried out for this model compared to the model containing hour as the only covariate. This was done to verify whether the day was contributing significantly more information than the hour alone.

Out-of-sample likelihood

The accuracy of the trained models could then be evaluated by calculating the out-of-sample Poisson loglikelihood of the datapoints in each of the withheld folds from the cross-validation using Equation 1, given the corresponding model trained on the other nine folds. This means that for each participant, ten log-likelihoods were calculated. The mean log-likelihood was then calculated for each participant to assist comparison between model types.

Simulation and visualisation

The trained models could then be used to simulate events occurring during the 'missing periods', by simulating point processes during a time period matching the withheld fold in terms of length and covariate values using NHPoisson's 'simNHP' function. λ was calculated according to Equation 2.

Hidden Markov Model comparisons

The simulated time series for each channel (i.e. phone usage, communication app usage, social media app usage) were then used to train two-state HMMs following the method in [26] using the depmixS4 package,[39] with the one-hot encoded hour of the day included in each model as a covariate on the transition probabilities. Two HMMs concerning imputed data were considered: an HMM trained on the

original imputed time series for each channel, and an HMM trained on imputed time series that were constrained by the imputed values in the phone usage channel. As it is not possible to have communication or social media app usage without phone usage, for the timepoints where there was no phone usage activity we set the values of the other channels to no activity as well. We compared each of these HMMs to the HMM trained on the corresponding 'real' data to investigate whether the HMM properties, i.e. the emission, transition and starting probabilities, were preserved.

Imputation application to real missingness

We then applied the imputation method to a combined dataset used in a prior digital phenotyping study and replicated the key findings. In the original study missingness was not imputed, rather the HMM handled missingness by omitting missing values from the log-likelihood calculations.

In the current study, we used the available data per participant to train an individualised non-homogeneous PPPM with one-hot encoded hour as a covariate. For participants whose model failed to converge, the 'optim' function with the 'BFGS' method was used to minimise the negative of the Poisson log-likelihood, in place of the default function. We then simulated their time series using this model and, where missingness occurred, inserted the values provided by their PPPM to complete their time series. For participants who had a channel with less than 3 activity points, we did not carry out imputation for that channel due to the limited amount of data available for training. This threshold was arbitrarily selected as for some channels datapoints could be very sparse, e.g. for people who do not use social media. Additionally, if a participant had more than 85% of their time series missing then no imputation was carried out as missingness was deemed too high for a representative model to be learnt. Given the study duration, this threshold required almost a week's worth of data per participant.

We then trained a two-state HMM on the completed time series as per the method in [26]. We chose to train the model on all participants rather than the training split used in [26] so that we could investigate the effect of imputation on the HMM for as many participants as possible, and therefore compared our results to the supplementary results of this paper that did not include a training/validation split. Note that the results from this original supplementary model were consistent with the results from the original main reported model.

We reported results from the constrained imputation method, where any imputed activity in other channels was removed if the imputed phone usage showed no activity. However, we did additionally investigate the unconstrained version where all channels have their original imputation values, and the results remained similar and significant (see Table S2 in the Supplementary Materials).

Acknowledgements

This study was funded by the European Research Council (consolidator grant 101001118). The SMARD study was funded by the Hersenstichting (Dutch Brain Foundation) (HA2015.01.07). The Dutch Brain Research Registry (Hersenonderzoek.nl) is supported by ZonMw-Memorabel (project no 73305095003), Alzheimer Nederland, Amsterdam Neuroscience, and Hersenstichting. The PRISM project has received funding from the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement 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.

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. All other authors declare no conflicts of interest.

Data availability

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

References

1. Jongs N, Jagesar R, van Haren NEM, et al. A framework for assessing neuropsychiatric phenotypes by using smartphone-based location data. Translational psychiatry. 2020;10(1):211. doi:https://doi.org/10.1038/s41398-020-00893-4
2. Jagesar RR, Roozen MC, van der Heijden I, et al. Digital phenotyping and the COVID-19 pandemic: Capturing behavioral change in patients with psychiatric disorders. European Neuropsychopharmacology. 1 2021;42:115-120. doi:https://doi.org/10.1016/j.euroneuro.2020.11.012
3. Ranjan Y, Rashid Z, Stewart C, 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. 2019;7(8):e11734. doi:https://doi.org/10.2196/11734
4. Jagesar RR, Vorstman JA, Kas MJ. Requirements and operational guidelines for secure and sustainable digital phenotyping: Design and development study. Journal of Medical Internet Research. 4 2021;23doi:https://doi.org/10.2196/20996
5. Bai R, Xiao L, Guo Y, 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. 2021;9(3):e24365. doi:https://doi.org/10.2196/24365
6. Saeb S, Zhang M, Karr CJ, et al. Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: an exploratory study. Journal of medical Internet research. 2015;17(7):e4273. doi:https://doi.org/10.2196/jmir.4273
7. Ware S, Yue C, Morillo R, et al. Predicting depressive symptoms using smartphone data. Smart Health. 2020;15:100093. doi:https://doi.org/10.1016/j.smhl.2019.100093
8. Vesel C, Rashidisabet H, Zulueta J, 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. 2020;27(7):1007-1018. doi:https://doi.org/10.1093/jamia/ocaa057
9. Benoit J, Onyeaka H, Keshavan M, Torous J. Systematic review of digital phenotyping and machine learning in psychosis spectrum illnesses. Harvard Review of Psychiatry. 2020;28(5):296-304. doi:10.1097/HRP.0000000000000268
10. Leaning IE, Ikani N, Savage HS, et al. From smartphone data to clinically relevant predictions: A systematic review of digital phenotyping methods in depression. Neuroscience & Biobehavioral Reviews. 2024:105541. doi:https://doi.org/10.1016/j.neubiorev.2024.105541
11. Cho C-H, Lee T, Kim M-G, In HP, Kim L, Lee H-J. Mood prediction of patients with mood disorders by machine learning using passive digital phenotypes based on the circadian rhythm: prospective observational cohort study. Journal of medical Internet research. 2019;21(4):e11029. doi:https://doi.org/10.2196/11029
12. Kathan A, Harrer M, Küster L, et al. Personalised depression forecasting using mobile sensor data and ecological momentary assessment. Frontiers in digital health. 2022;4:964582. doi:https://doi.org/10.3389/fdgth.2022.964582
13. Pellegrini AM, Huang EJ, Staples PC, et al. Estimating longitudinal depressive symptoms from smartphone data in a transdiagnostic cohort. Brain and Behavior. 2022;12(2):e02077. doi:https://doi.org/10.1002/brb3.2077
14. Zhang Y, Folarin AA, Sun S, et al. Predicting depressive symptom severity through individuals' nearby bluetooth device count data collected by mobile phones: preliminary longitudinal study. JMIR mHealth and uHealth. 2021;9(7):e29840. doi:https://doi.org/10.2196/29840
15. Lee H-J, Cho C-H, Lee T, et al. Prediction of impending mood episode recurrence using real-time digital phenotypes in major depression and bipolar disorders in South Korea: a prospective nationwide

cohort study. Psychological medicine. 2023;53(12):5636-5644. doi:https://doi.org/10.1017/S0033291722002847
16. Van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate imputation by chained equations in R. Journal of statistical software. 2011;45:1-67. doi:https://doi.org/10.18637/jss.v045.i03
17. Liu G, Onnela J-P. Bidirectional imputation of spatial GPS trajectories with missingness using sparse online Gaussian process. Journal of the American Medical Informatics Association. 2021;28(8):1777-1784. doi:https://doi.org/10.1093/jamia/ocab069
18. Saha K, Reddy MD, das Swain V, et al. Imputing missing social media data stream in multisensor studies of human behavior. IEEE; 2019:178-184. doi:10.1109/ACII.2019.8925479
19. Tackney MS, Cook DG, Stahl D, Ismail K, Williamson E, Carpenter J. A framework for handling missing accelerometer outcome data in trials. Trials. 2021;22(1):379. doi:https://doi.org/10.1186/s13063-021-05284-8
20. Tackney MS, Williamson E, Cook DG, Limb E, Harris T, Carpenter J. Multiple imputation approaches for epoch-level accelerometer data in trials. Statistical Methods in Medical Research. 2023;32(10):1936-1960. doi:https://doi.org/10.1177/09622802231188518
21. Lee MP, Hoang K, Park S, et al. Imputing missing sleep data from wearables with neural networks in real-world settings. Sleep. 2024;47(1):zsad266. doi:https://doi.org/10.1093/sleep/zsad266
22. Chakrabarti S, Biswas N, Karnani K, et al. Binned data provide better imputation of missing time series data from wearables. Sensors. 2023;23(3):1454. doi:https://doi.org/10.3390/s23031454
23. Sasu G-V, Ciubotaru B-I, Goga N, Vasilățeanu A. Addressing Missing Data Challenges in Geriatric Health Monitoring: A Study of Statistical and Machine Learning Imputation Methods. Sensors. 2025;25(3):614. doi:https://doi.org/10.3390/s25030614
24. Behapp. Accessed Jul 3 2024, www.behapp.com
25. Ikani N, Tyborowska A, Kohn N, et al. Smartphone-based Monitoring and cognition Modification Against Recurrence of Depression (SMARD): An RCT of Memory Bias Modification Training vs. Cognitive Control Training vs. Attention Bias Modification Training in remitted recurrently depressed patients with 1.5 year follow-up. BMC psychiatry. 2025;25(1):445. doi:https://doi.org/10.1186/s12888-025-06860-x
26. Leaning IE, Costanzo A, Jagesar R, et al. Uncovering Social States in Healthy and Clinical Populations Using Digital Phenotyping and Hidden Markov Models: Observational Study. Original Paper. J Med Internet Res. 2025;27:e64007. doi:10.2196/64007
27. Muurling M, Reus LM, de Boer C, et al. Assessment of Social Behavior Using a Passive Monitoring App in Cognitively Normal and Cognitively Impaired Older Adults: Observational Study. JMIR Aging. April 2022;5doi:https://doi.org/10.2196/33856
28. First MB, Gibbon M, Spitzer RL, Williams JBW. User Guide for the Structured Clinical Interview for DSM-IV Axis-1 Disorders. Washington DC: American Psychiatric Association. 1996;
29. Hamilton M. Development of a rating scale for primary depressive illness. British journal of social and clinical psychology. 1967;6(4):278-296. doi:https://doi.org/10.1111/j.2044-8260.1967.tb00530.x
30. Kas MJH, Jongs N, Mennes M, et al. Digital behavioural signatures reveal trans-diagnostic clusters of Schizophrenia and Alzheimer's disease patients. European Neuropsychopharmacology. 2024;78:3-12. doi:https://doi.org/10.1016/j.euroneuro.2023.09.010
31. Bilderbeck AC, Penninx BW, Arango 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. 2019;97:87-93. doi:https://doi.org/10.1016/j.neubiorev.2018.06.019
32. Kas MJ, Penninx B, Sommer B, Serretti A, Arango C, Marston H. A quantitative approach to neuropsychiatry: the why and the how. Neuroscience & Biobehavioral Reviews. 2019;97:3-9. doi:https://doi.org/10.1016/j.neubiorev.2017.12.008
33. Birchwood M, Smith JO, Cochrane R, Wetton S, Copestake SONJA. 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. 1990;157(6):853-859. doi:https://doi.org/10.1192/bjp.157.6.853
34. de Jong-Gierveld J. Developing and testing a model of loneliness. Journal of personality and social psychology. 1987;53(1):119. doi:https://doi.org/10.1037/0022-3514.53.1.119
35. Kay SR, Fiszbein A, Opler LA. The positive and negative syndrome scale (PANSS) for schizophrenia. Schizophrenia bulletin. 1987;13(2):261-276. doi:https://doi.org/10.1093/schbul/13.2.261
36. Folstein MF, Folstein SE, McHugh PR. "Mini-mental state": a practical method for grading the cognitive state of patients for the clinician. Journal of psychiatric research. 1975;12(3):189-198. doi:https://doi.org/10.1016/0022-3956(75)90026-6
37. Mulder T, Jagesar RR, Klingenberg AM, Bonnici JPM, Kas MJ. New European privacy regulation: Assessing the impact for digital medicine innovations. European Psychiatry. 2018;54:57-58. doi:https://psycnet.apa.org/doi/10.1016/j.eurpsy.2018.07.003
38. Cebrián AC, Abaurrea J, Asín J. NHPoisson: An R package for fitting and validating nonhomogeneous Poisson processes. Journal of Statistical Software. 2015;64:1-25. doi:https://doi.org/10.18637/jss.v064.i06
39. Visser I, Speekenbrink M. depmixS4: an R package for hidden Markov models. Journal of statistical Software. 2010;36:1-21. doi:https://doi.org/10.18637/jss.v036.i07