Artificial Intelligence In Medicine 142 (2023) 102575

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

journal homepage: www.elsevier.com/locate/artmed

Prediction meets time series with gaps: User clusters with specific usage behavior patterns

Miro Schleicher [a,*], Vishnu Unnikrishnan [a], Rüdiger Pryss [b], Johannes Schobel [c], Winfried Schlee [d], Myra Spiliopoulou [a,*]

[a] Knowledge Management & Discovery Lab, Otto-von-Guericke-University Magdeburg, Magdeburg, Germany
[b] Institute of Clinical Epidemiology and Biometry, University of Würzburg, Würzburg, Germany
[c] Institute DigiHealth, Neu-Ulm University of Applied Sciences, Neu-Ulm, Germany
[d] Eastern Switzerland University of Applied Sciences, St. Gallen, Switzerland

ARTICLE INFO

Keywords:
Time series with gaps
Adherence
Law of attrition
Chronic diseases
mHealth

ABSTRACT

With mHealth apps, data can be recorded in real life, which makes them useful, for example, as an accompanying tool in treatments. However, such datasets, especially those based on apps with usage on a voluntary basis, are often affected by fluctuating engagement and by high user dropout rates. This makes it difficult to exploit the data using machine learning techniques and raises the question of whether users have stopped using the app. In this extended paper, we present a method to identify phases with varying dropout rates in a dataset and predict for each. We also present an approach to predict what period of inactivity can be expected for a user in the current state. We use change point detection to identify the phases, show how to deal with uneven misaligned time series and predict the user's phase using time series classification. In addition, we examine how the evolution of adherence develops in individual clusters of individuals. We evaluated our method on the data of an mHealth app for tinnitus, and show that our approach is appropriate for the study of adherence in datasets with uneven, unaligned time series of different lengths and with missing values.

1. Introduction

The option of recording data in real life is a major advantage of mHealth apps. They are therefore particularly helpful when accompanying treatments. Notwithstanding the potential benefits to all involved, the use of such apps requires the willingness and discipline of the individuals involved to participate consistently. Data coming from such sources are often affected by fluctuating engagement and by high dropout rates. Application of machine learning techniques to such datasets are therefore confronted with these challenges. In concrete terms, a number of problems arise which complicate the handling of the data. First, the high dropout rates that are the subject of the *science of attrition* initiated by Eysenbach [1]. Second, the fluctuating engagement during use. This creates gaps in the data of varying size (*missing data*). Third, the sampling of the datasets. There is rarely even spacing between surveys or an equal number of observations. Engagement with the app is beneficial for all parties involved, so it is in the app provider's best interest to assess whether a gap in the data means abandonment or whether the person is likely to return. Therefore, we present a method to identify phases with different dropout rates according to Eysenbach and make a prediction for each phase. We also present an approach to predict what period of inactivity to expect for a user in the current state. Therefore, we raise the question to what extent it is possible to identify and predict phases with different dropout rates in these datasets, as well as to predict the duration of inactivity that can be expected from a user in a current state. We use *Change Point Detection* (CPD) to identify phases, show how to deal with uneven, misaligned time series, and predict the user's phase using time series classification. Beyond that, we learn about clusters of app users based on certain key characteristics and observe their evolution in the context of the identified phases. We evaluated our method on data from an mHealth app for tinnitus and show that our approach is suitable for studying adherence in datasets with uneven, unaligned time series of different lengths and with missing values. The scope of the paper refers to the presentation of a possible workflow and the demonstration of the application on a real use case. The focus is on the methods and less on the insights gained from the data set.

The paper is organized as follows: Section 2 presents the related works; Section 3 encompasses the methods used to solve the problems and Section 4 describes the material with which we tested our

* Corresponding authors.
E-mail addresses: miro.schleicher@ovgu.de (M. Schleicher), myra@ovgu.de (M. Spiliopoulou).

https://doi.org/10.1016/j.artmed.2023.102575
Received 13 January 2023; Received in revised form 25 March 2023; Accepted 27 April 2023
Available online 2 May 2023

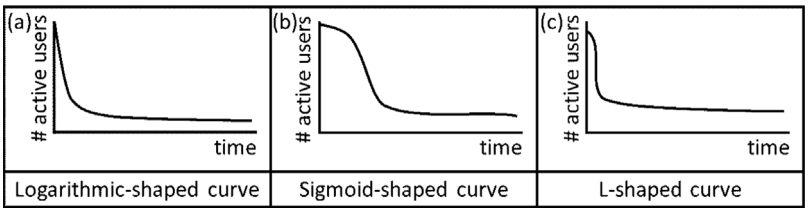

Fig. 1. The three types of attrition curve identified in [1], here plotted as prototypical curves.

method and shows the results obtained. Finally, Section 5, draws the conclusions and provides directions for future work.

2. Related work

Related to our approach is research in the *science of attrition*, advances on adherence/compliance modeling and monitoring as well as investigations in dealing with time series with gaps. *Engagement* in the use of self-monitoring apps is often referred to as *adherence* or *compliance*. The WHO has summarized adherence as: "the extent to which a person's behavior [...] corresponds with agreed recommendations [...]" [2]. They differentiate between adherence and compliance by the agreement of the patient to the recommendations (adherence) [2]. In our study, we investigate 'adherence' in the context of interaction with an mHealth app, and define as 'dropouts' the persons that give up and stop future interactions. Then, we investigate to what extent Eysenbach's *law of attrition* [1] for survey data [3,4] and for longitudinal experimental data [1] can also be applied on our longitudinal, observational data.

In [1], Eysenbach investigated how the percentage of study participants still engaged in a (medical) study changes from the beginning to the end. He collected data from several studies, and he found that attrition in some studies follows a sigmoid-shaped curve characterized by an initial plateau of high participation, while attrition in other studies follows more of an 'L-shaped curve'; logarithmic-shaped curves have also been identified. These three types of curve are plotted in Fig. 1 for easier juxtaposition. These attrition curves though, were plotted for studies where a participant who gives up does not come back. Eysenbach mainly describes 3 phases [1], namely the *curiosity plateau* (Phase I) as the initial phase where the user are interested in exploring a new technology, followed by the *attrition phase* (Phase II) where the users start to reject the usage and finally, the *stable use phase* (Phase III) where only the 'hardcore users' remain, which will continue to use the application for a long time [1].

Cismondi et al. investigate methods on medical data to deal with the problem and propose alignment methods such as gridding and templating [5]. However, the principle cannot be applied to the data of this paper, since the time series of a user are not collected separately but by the same questionnaire and therefore already have the same time stamp. Since the generation of the time series of the individual users differs substantially in some cases, the principle can only be applied here to a limited extent. We proposed a model of adherence based on such data in our previous work [6], but we did not attempt to predict adherence specifically. Such predictions can be found for specific applications. For example, Williams-Kerver et al. predicted adherence in eating disorder based on data with gaps, but focused on person-level characteristics, such as gender, rather than the data records themselves [7]. We presented in [8] a recommender method based on a matrix factorization approach also on longitudinal, observational data to predict time periods without data to be expected based on uninterrupted sequences of input data right before them. This is another more complex approach than the 1-Nearest Neighbor classification presented in this paper. The advantage of the latter approach is its well acknowledged performance [9] and the comparatively easier applicability. Our paper [10] also deals with the predictability of adherence. It presents four methods: a shapelet-based predictor, a dictionary-based predictor, one based on matrix profiles, and a windows-based approach. The goal is to predict, with a given input of data for a user-selected point in time in the future, whether a person is adherent at that moment or not. However, the modeling of adherence and the information used with it differed from the approach used in this paper. For example, in [10] a general adherence level is introduced, which, for example, takes short-term non-adherence less into account if the person was previously adherent for a long time than for very fluctuating persons.

Furthermore, this work must be distinguished not only by the data used (survey data [3,4] or longitudinal experimental data [1] vs. longitudinal observational data) but also by the source or intention of the data source. For example, the retrieval of health status questionnaires from voluntarily used self-monitoring apps differs from that of, e.g., an app included in a randomized controlled trial (RCT), for which there may be usage protocols and thus other commitment relationships of the users (e.g., in [11]). This also applies to other apps that, for example, monitor the intake of medication or the adherence to a therapy (e.g., [12]) and thus also have other adherence requirements, or to fitness apps (e.g., [13]), which in turn have other adherence barriers due to increased requirements (effort), to give just two examples.

3. Materials and methods

In this section, the dataset for the evaluation is presented at the beginning. Subsequently, the approach is described and its most important steps are addressed in more detail in separate subsections.

3.1. Dataset

The evaluation is performed on a mHealth dataset of the TrackYourTinnitus (TYT) [14] self-monitoring app, dedicated to research on tinnitus and to help users understand their manifestation of the disorder. Tinnitus is a complex chronic disorder that has no uniform way of manifestation and generation [15]. The initial dataset contains 3177 users with a gender distribution of 1028 females, 2097 males and 52 users with no specified gender. The observational period is from 2014-04-10 till 2022-01-17. The mean age is 45.50 years with a standard deviation (STD) of 13.20. The mean age of tinnitus onset is 36.02 years (STD:14.94). When the app randomly sends a request, the users should answer 8 questions if possible. Each of these 8 questions (c.f. [6] Table 1) will form separate time series. How often (or if at all) each individual user adds data points to the time series per day is up to them and cannot be generalized. Also, whether they answer one, several, or all questions in a session. In order to demonstrate our approach, we will focus in this paper the question "How stressful is the tinnitus right now?" ("distress"), but it would be applicable on the other questions as well.

3.2. Modeling of the time series of the users

The starting point are the time series generated by the user. A time series consists of the transmitted answers (range [0, 1]) to a question. The question can be asked randomly 1:n times per day (by user choice) and can also be refused by the user. In this work we refer to a single question and thus to uni-variate time series.

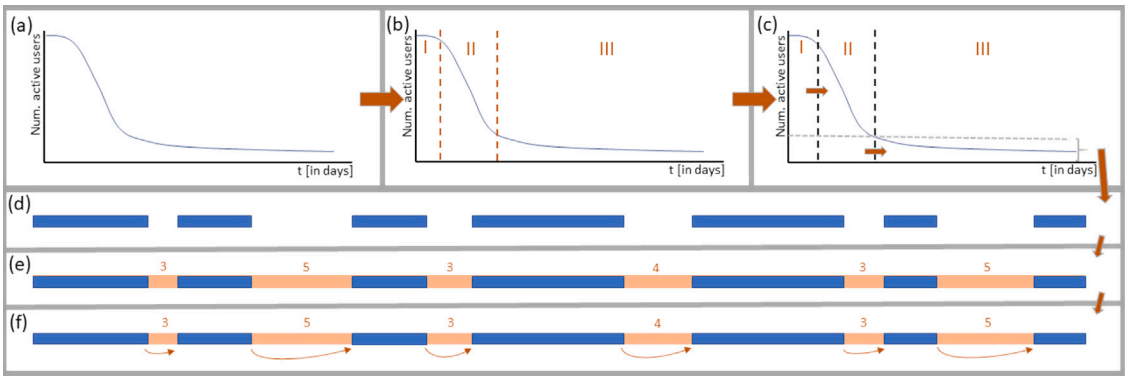

Fig. 2. Big Picture. (a) Attrition curve, (b) Phases of attrition, (c) Predicting the phases, (d) A time series represented as sequences (symbolized example), (e) Sequences with labeled gaps, (f) Predicting the return.

3.3. Modeling three phases in the time series of the users

The first step is to identify the phases with varying dropout rates (c.f. Fig. 2(a)), since according to Eysenbach these represent specific stages of interaction. All three phases are to be modeled, but we also check if indeed all phases are observable or if one phase has no data. Fig. 2(b) symbolizes the phases, which are snapshots and can change over time. Determining them in an automated manner might help to monitor the developments of the dataset. We apply CPD methods to determine the change in dropout rates and use them to determine the phases (c.f. transition Fig. 2(a) to (b)).

Using this information, we can now try to predict whether a user will reach the next phase (from Phase I to II and from I+II to Phase III) as symbolized in Fig. 2(c). A prediction from Phase I to III does not seem reasonable, as the transition from Phase I to Phase II is already characterized by a high dropout rate. Therefore the information in II may be crucial in predicting III.

Individuals in Phase III are considered stable in their continued use, but their usage patterns may fluctuate. Fig. 2(d) shows an example of a single user's time series consisting of 7 sequences (blue) with responses and 6 gaps (white). Therefore, the next step is to predict when individuals most likely will return after a break. The basis here is no longer all data before the gap in question, but only the sequence between the last gap and the ones to be predicted. Thus, the approach respects the high variability and the sparsity problem. Each contiguous sequence of data (at least two consecutive days) is assigned the value of the gap until the next entry, providing evidence of the return. The duration of a gap is counted in days. Fig. 2(e) symbolizes these gaps between the sequences as orange sections and the numbers above them represent the number of days without data. This cannot apply to a user's last sequence, which hides the information if the user will contribute again. Possible reasons could be a dropout or the database has reached its cut-off date, to name just two. Sequences and gaps will have a high variability in length. To facilitate classification, the gaps, symbolized in Fig. 2(e), can now be grouped together (binning) according to their size in order to form categories. Each category is intended to represent an interval of absence in which similar users have returned. The mean of the interval (in days) corresponds to the expected value of the class and the minimum and maximum values indicate a kind of uncertainty range. Fig. 2(f) represents the classification of these target classes by inferring the duration of the gaps from the sequence, represented as arrows from one sequence to the beginning of the next.

3.4. Identifying phases of attrition

A potential target for CPD are Eysenbach's [1] proposed *attrition curves*. In the self-monitoring context a user can return even after a very long time. As a consequence, the true number of enrolled users must be assumed as unknown. Therefore, we created our curve by the number of contributing users over time. Where 'time' is not measured in dates, moreover in subjective days of app usage. The first day (day 1) is the day of the first submitted record. From that day on, the days are counted as usual from 0:00 a.m. on. On each of these days of usage some users might contribute and some might pause in varying combinations. But some users might start at some point to stop their contribution. As a result the number of contributing persons will decrease over the time of aligned days. In order to make this approach applicable to multiple sample of users and applications we decided to search for all 3 phases. Based on the underlying data, it is possible that the initial phase comprises only one day, namely day 1, and the curve thus corresponds more to an L-shaped attrition curve.

We applied 3 change point detection methods: *Linearly penalized segmentation* (PELT) [16], *Dynamic programming* (Dynp) [16] and *Bottom-up segmentation* (BottomUp) [16–18] from the python package 'rupture' [16]. To select the best method for the data, the selection must be made initially after visual inspection.

3.5. Predicting attrition for each user

For predicting whether a user will contribute in the next phase, the algorithm *XGBoost* from the Python package 'XGBoost' [19] was selected. It iteratively combines different models in the eponymous boosting procedure to reduce the errors of those already implemented. The algorithm was chosen for its broad applicability and excellent performance. Since this is a binary classification, a 10-fold cross-validation (CV) with *Accuracy* (Acc) was chosen to evaluate the results.

To clarify, the prediction target is the *adherence class*, which refers to the information whether a user is active in a found phase. The *adherence in the variable* refers to whether the user has answered the question of the app on the corresponding day or not. The adherence in the variable is thus decisive for the assignment of the adherence label in the phases (respectively with the class 'yes' and 'no'). If a user is not active in Phase II, but is active in later phases, the person is labeled as not adherent in Phase II, but as adherent in Phase III.

3.6. Learning clusters of users

After determining the phases with varying dropout rates and making predictions for the totality of users, we learn clusters of users on selected key characteristics. We contrast how the percentage distribution of clusters change in Phase II & III compared to Phase I under the condition that users are adherent or not. This approach makes it possible to trace the evolution of clusters with a focus on adherence and to identify and describe groups of users that stand out in their usage behavior.

To explore the possibilities of the clusters, we use different features that can characterize a tinnitus patient. For this purpose, the TYT dataset provides us with the *Mini Tinnitus Questionnaire* (Mini-TQ) [20]

and the *Tinnitus Sample Case History Questionnaire* (TSCHQ) [21]. The Mini-TQ measures tinnitus-related psychological distress [20]. This questionnaire is an abbreviated form of the Tinnitus Questionnaire (TQ) [22,23] which, among others, is considered an "essential part of patient assessment" [21] and is valued as an "outcome measurement" [21]. In [20] it is shown that the reduced version to 12 questions and the sum score calculated from these are comparable to the full version and "no recognizable psychometric disadvantages" [20] exist to the TQ. Therefore, we use the Mini-TQ sum score as a feature for clustering.

The other features are selected from TSCHQ from the areas "Background", "Tinnitus history", "Modifying influences" and "Related conditions" [21]. Case history questionnaires provide the experts with information on descriptive characteristics of the patients' tinnitus as well as the related conditions [21].

The selected items of the TSCHQ are [24]:

Background
- Age at visit (years)*
- Gender (female/male)*

Tinnitus history
- Age at tinnitus onset (years) (author remark: modified from 'time in month')
- Onset related events (change in hearing/ stress/loud blast of sound/head trauma/whiplash/others)
- Subjective tinnitus loudness (0–100)*

Modifying influences
- Influence by stress (worsen/reduces/no effect)

Related conditions
- Hearing impairment (no/yes)*
- Noise intolerance (never/rarely/ sometime/usually/always)
- Headache (no/yes)*
- Temporomandibular joint complaints (no/yes)
- Neck pain (no/yes)
- Other pain (no/yes)
- Psychiatric problems (no/yes)

The variables were chosen to be drawn from all areas of the TSCHQ and to be among those considered "essential" [21] (marked with *) as well as "highly desirable" [21] (the rest) information. Other combinations and numbers of items are also possible. In the proposed selection it was important that the items are understandable for the study of the clusters, are composed of both numerical and categorical variables and, in addition, present a particular adherence behavior in the concrete evaluation dataset.

Clustering was performed using the HDBSCAN algorithm [25,26]. The approach has the advantage of using only one parameter (the minimum size of the clusters), which facilitates tuning. At the same time, the density-based approach offers the possibility to ignore records that do not fit into a certain cluster as "noise". Especially in very heterogeneous datasets, clusters can benefit from this property.

The distance between two instances was calculated using "Jaccard" distance. Since questionnaires must expect that many questions may not be answered, a method is needed that can also handle these missing values. This is different from e.g. Euclidean distance, where these variables have to be either ignored or changed, which can lead to distortions, e.g. when missing values are counted as '−1' in the distance calculation. Furthermore, the information whether a variable was answered or not is itself already information that can be worked with.

A challenge in tuning the parameter of HDBSCAN is that a ground truth is missing and that not every instance is assigned to a cluster. This complicates the application of traditional methods for optimization. In the present use case, however, it is already interesting to show that stable clusters can be found in the context and the corresponding adherence patterns.

The evaluation of a selection of the clusters is done manually in this paper to explore and better assess the possibilities of clustering in the process.

3.7. Tuning the gap size for prediction

To create the necessary preconditions for the classification of the sequences, restrictions must be made. User with just a single day and therefore, with just a single day sequence must be excluded. This is also true for users with multiple days but no second sequence, since the return after the gap cannot be verified. If a uni-variate sequence in a multivariate sequence has a missing value on one day, this sequence must be removed (although imputation can be explored) Finally, gaps of a certain length might not worth considering. Depending on the use case of the app, such sequences are not very informative, because the user might try a restart or test a new version of the app after this very long pause. They are sequences with return, but are basically 'hidden dropouts'. We considered three strategies for data binning, namely (a) building equisized intervals, (b) building intervals on frequency and (c) identifying 'natural' groups with the *Fischer–Jenks algorithm* [27] (implementation: *jenkspy* package). Each algorithm delivers a different number of bins of different sizes, which must be categorized manually.

The prediction of the class of a sequence is done using a 1-NN classifier with Dynamic Time Warping (DTW), one of the best performing approaches according to [9]. Since the binning will lead to multiple classes with uneven distribution, a stratified 10-fold CV is selected for evaluation with Acc.

4. Results and discussion

The following sections describe the results for identifying the attrition phases, followed by the predictions of the phases for each user, and some details on tuning the gap size parameter. We close with a discussion of the results.

4.1. Identifying phases of attrition

Fig. 3 depicts the attrition curves, computed as described in Section 3.4. We see that on day 1, all 3177 users were active. On day 2, there were 1284 (40.41%), on day 3 1011 (31.82%) and on day 14 only 481 (15.14%) users. We also see that the curve flattens after day 11 and there is practically no change after day 12. Therefore, we focus on the first $12 + 2 = 14$ days, i.e. we take two more days into account for our attrition study. Although it would be technically possible to consider more days, we believe that the caregiver or study coordinator should focus on early signals of attrition at the beginning of the study; this is an additional motivation for considering only the first two days after the flattening.

Visual inspection of the graph and the rapid decline in numbers suggest an L-shaped progression rather than the sigmoidal progression shown in [1]. When running the 3 CPD algorithms (PELT, BottomUp, Dynp) on the data shown in Fig. 3, quite different results were obtained. All 3 algorithms were expected to find 2 change points (or 3 phases) in the first 14 days after visual inspection of the curve, but were set to freely find the change points. PELT detected only a single change point on day 2, while BottomUp found 2 points. The first on day 5 and the second on day 14. Dynp detected the first change point on day 2, just like PELT. The second point on day 5, which is the same result as BottomUp. And another one on day 14. In experiments with shorter or longer time series, the algorithms BottomUp and Dynp selected day 14, i.e. the last day, so this last change point should be discarded. Hence, PELT, BottomUp and Dynp, taken together, identified change points

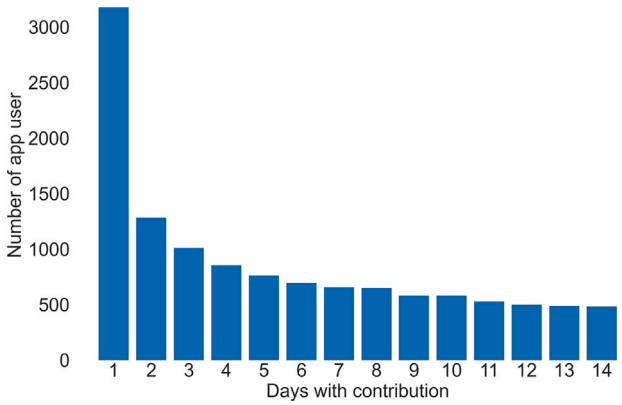

Fig. 3. The figure illustrates the number of users contributing data on Day 1 through Day 14. Day 1 is the first day of use of the app for each individual user. Not every user who submitted data on Day 1 is present in all other days.

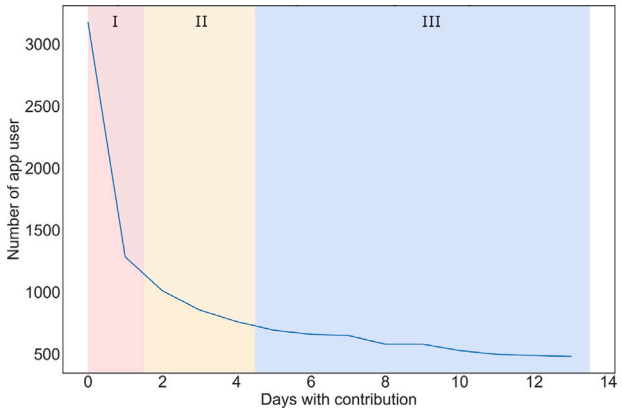

Fig. 4. This figure shows the result of the dynamic programming search method for change points (Dynp). The color change symbolizes a different area marked by the calculated change points on day 2 and day 5. Phase I (red) - II (yellow) - III (blue). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

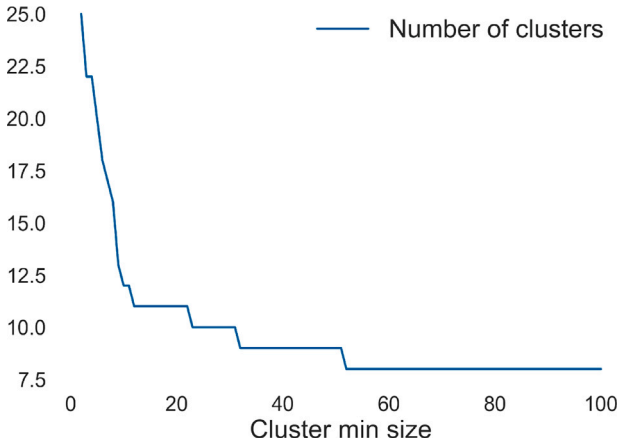

Fig. 5. The figure shows the number of clusters for different settings of the minimum cluster size (cluster min size) for the HDBSCAN algorithm. The number of clusters stabilizes at a minimum size of 52 instances to 7 clusters + noise.

at day 2 (two of the three algorithms) and day 5 (two algorithms). These two changepoints indicate that there are indeed three phases, in agreement with the law of attrition. However, the change from Phase II to Phase III (at day 5) is not so prominent, as can also be seen on Fig. 4.

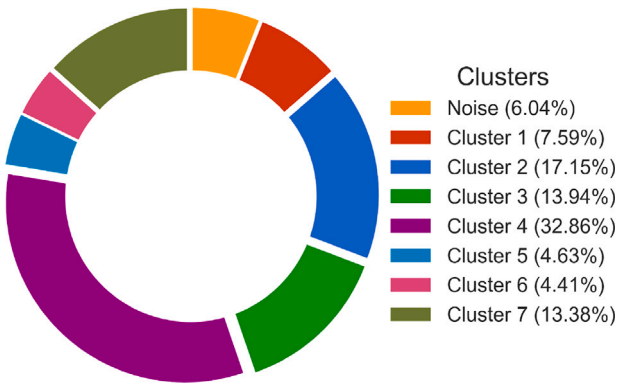

Fig. 6. This figure illustrates the percentage distribution of clusters and the proportion of instances considered "noise" for HDBSCAN with a minimum cluster size of 52 instances.

In that figure, the red area (leftmost) refers to Phase I, the yellow area (middle) to Phase II and the blue area to Phase III.

4.2. Predicting attrition for each user

Since Dynp found both changepoints in the time series, we used it to assign labels to users. Anyone who contributed on day 2 or later was assigned a Yes class for Phase II, and everyone else was assigned a No class. A similar procedure was performed with the threshold for Phase III on day 5. As input for the classifiers, we chose the time series 'distress' because this variable reflects best how the users feel about their tinnitus [15] and because it is likely to be associated to attrition. We excluded 99 users who did not enter distress values on day 1, whereupon we retained 3078 users with an entry in day 1 (i.e. before the change point at day 2). These users we used for prediction. The classes are slightly imbalanced, with $n = 1811$ (58.84%) for Yes and $n = 1267$ (41.16%) for No. Since the users have substantial differences in the data contributed on the days and also at very different times, the median per day was chosen to be representative, so that each user has only one value per day. So in this specific case, the algorithm had to try to make a prediction with one value per 3078 users.

One value per day, making the classification task very difficult. We used XGBoost [19] with 10-fold cross-validation and achieved a mean accuracy of 54.52% (SDT:3.68%). The low accuracy reflects the difficulty of the task. For the prediction of Phase III, a much higher quality was achieved, namely 76.93% (SDT:1.93%). The class distribution is Yes: 1435 (45.17%) and No: 1742 (54.83%).

4.3. Learning clusters of users

After the predictions of the adherence classes, the clusters are now in focus. At the beginning, it is examined to what extent the proposed variables are filled in. The adherence per variable is: age: 96.85%, gender: 99.36%, onset age: 95.37%, Mini-TQ sum score 100.00%, initial onset: 98.11%, loudness: 82.18%, effect of stress: 74.28%, hearing impairment: 73.62%, noise induced pain: 73.62%, headaches: 73.62%, vertigo/dizziness: 73.59%, temporomandibular disorder: 73.62%, neck pain: 73.62%, other pain: 73.53% and psychiatric problems: 73.59%. The selected variables of TSCHQ's domains "modifying influences" and "related conditions" were not answered by more than a quarter of the users, while the participation in the other variables was high, with the exception of loudness and the effect of stress.

Since the HDBSCAN algorithm has just one parameter, the minimum cluster size, different settings were tested until a stable cluster size was found. The number of clusters stabilizes at a minimum size of 52 instances to 7 clusters + noise as illustrated in Fig. 5. The resulting cluster distribution is shown in Fig. 6 as a percentage as well as the

Table 1
This table represents the distribution of the clusters in the different adherence classes with respect to the phases. The total count of participants (#), the percentage (%) per class and phase and the tendency (T) of the cluster development given the percentage of the previous phase. The tendency is shown as an arrow. In Phase II the previous phase is Phase I (T_{p1}) and in Phase III it is the tendency from Phase II (T_{p2}).

Cluster	All		Adherent						Non-Adherent					
	Phase I		Phase II			Phase III			Phase II			Phase III		
	# (of 3177)	%	# (of 1848)	%	T_{p1}	# (of 1435)	%	T_{p2}	# (of 1329)	%	T_{p1}	# (of 1742)	%	T_{p2}
Cluster 1	241	7.59	123	6.66	↘	96	6.69	→	118	8.88	↗	145	8.32	↘
Cluster 2	545	17.15	358	19.37	↗	275	19.16	↘	187	14.07	↘	270	15.50	↗
Cluster 3	443	13.94	220	11.90	↘	179	12.47	↗	223	16.78	↗	264	15.15	↘
Cluster 4	1044	32.86	653	35.34	↗	511	35.61	↗	391	29.42	↘	533	30.60	↗
Cluster 5	147	4.63	84	4.55	↘	64	4.46	↘	63	4.74	↗	83	4.76	→
Cluster 6	140	4.41	64	3.46	↘	48	3.34	↘	76	5.72	↗	92	5.28	↘
Cluster 7	425	13.38	246	13.31	→	180	12.54	↘	179	13.47	↗	245	14.06	↗
Noise	192	6.04	100	5.41	↘	82	5.71	↗	92	6.92	↗	110	6.31	↘

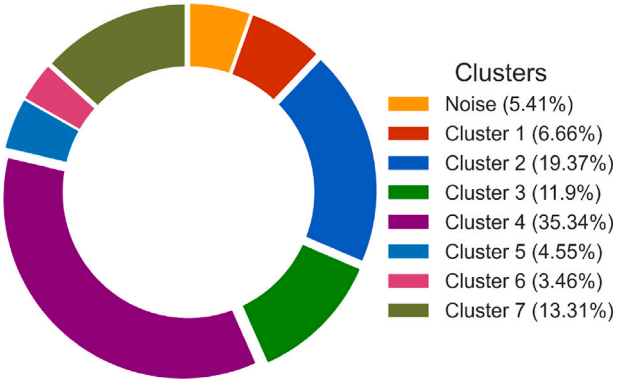

Fig. 7. The figure shows the cluster distribution in percent for all users who were adherents in Phase II. The plot corresponds to the values in Table 1 in the columns "Adherent - Phase II".

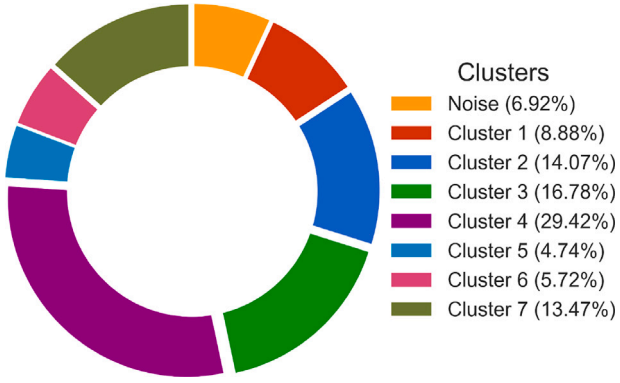

Fig. 9. The figure shows the cluster distribution in percent for all users who were non-adherents in Phase II. The plot corresponds to the values in Table 1 in the columns "Non-Adherent - Phase II".

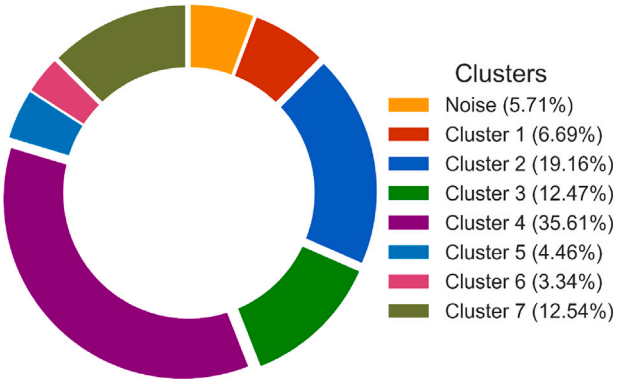

Fig. 8. The figure shows the cluster distribution in percent for all users who were adherents in Phase III. The plot corresponds to the values in Table 1 in the columns "Adherent - Phase III".

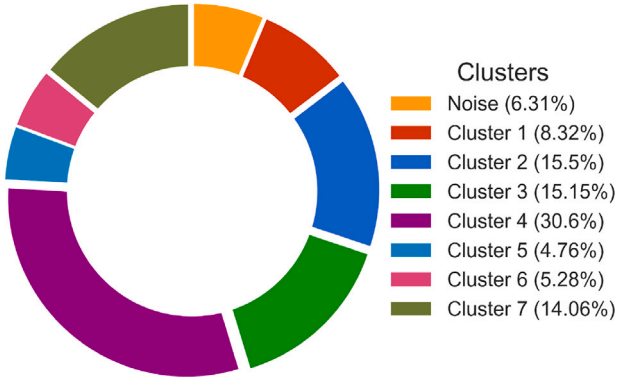

Fig. 10. The figure shows the cluster distribution in percent for all users who were non-adherents in Phase III. The plot corresponds to the values in Table 1 in the columns "Non-Adherent - Phase III".

fraction of instances labeled as "noise". Table 1 contains the same information for the Phase I columns plus the total number of cluster members. The amount of people marked as noise is only 6.04%.

Table 1 also distinguishes between a behavior that is adherent and one that is not. In these behavior phases, the composition of users by cluster is examined and the tendency of evolution of the percentage distribution compared to the previous phase is symbolized by arrows. Thus, the trend per cluster can be compared for both groups of users (adherent vs. non-adherent). A user who is adherent in Phase II but not in Phase III changes the group to non-adherent, but still counts to the same cluster. Figs. 7 & 8 illustrate the distribution for Phase II & III for adherent users. Figs. 9 & 10 do the same for non-adherent persons.

The "noise" cluster in Phase II almost splits into two halves in terms of absolute numbers (100 vs. 92). In percentage terms, the adherent group has less noise and the non-adherent group more. In Phase III, it increases again for the first group and decreases for the second.

Nevertheless, the final percentage for the adherence group is 5.71%, which is below the level of Phase I, and 6.31% for the non-Adherent group, which is above.

Cluster 4 is the largest cluster with 1044 members. Slightly more than one-third of the members are not adherent in Phase II. The proportion of the adherent group increases slightly with the phases and is above the level in Phase I. Among the second group, it first increases and then decreases, but always remains below the level in Phase I. But in both cases it remains the largest cluster.

Cluster 5 & 6 are the smallest clusters. Both clusters are characterized by the fact that their proportion in the adherent group continues to decrease. In the non-adherent group, the proportion first increases for both. However, cluster 5 then remains relatively stable, while the level of cluster 6 decreases.

Cluster 7 remains relatively stable in the first group in Phase II and then decreases slightly. In the non-adherent group, however, the percentage rises slightly above the initial level from phase to phase.

Using one of the clusters and the noise group as an example, the manual analysis of the clusters will now be demonstrated. Cluster 4 is the first example as the cluster with the most users. It is the largest cluster by far (32.86% of all users). The next largest is cluster 2 with 545 people. It is also worth noting that the algorithm assigned only male users to this cluster. In addition, the people are the oldest, with a mean age of 48.75 years (STD: ±13.13 years) compared to the average of the other clusters (without noise), with a mean age of 43.74 years (mean STD: ±12.82 years). The age of onset is slightly above the mean of the other clusters, with a mean onset at 37.58 years (STD: ±14.53 years) to an average of 36.00 ± 14.15 years. The Mini-TQ sum score is slightly below the average of the other clusters, with 13.84 ± 5.87 to an average of 14.09 ± 5.72. In the Phase II adherence group, the mean age is noticeably higher at 50.05 ± 12.54 compared to the average of the other clusters (45.09 ± 13.02), and the age at onset is slightly higher at 39.27 ± 13.95 to 37.04 ± 14.72. The Mini-TQ sum score comparable with 13.90 ± 5.87 to 13.98 ± 5.63. In Phase III, the mean age is comparable to Phase II with 50.80 ± 12.21 (average other clusters: 45.47 ± 13.05). The onset age also keeps its level 39.52 ± 13.58 (average other clusters: 37.10±14.51). And so does the Mini-TQ sum score at 13.74±5.89 (average other clusters: 13.91 ± 5.54). In the non-adherence Phase II group, mean age is lower in comparison to Phase I with 46.52 ± 13.81 years, but the average of the other clusters: 42.15 ± 12.45 is noticeably lower, also age at onset is lower with 34.67 ± 15.06 (average other clusters: 34.71 ± 13.43) and Mini-TQ sum score at 13.73 ± 5.88 (average other clusters: 14.23±5.81) shows no major differences from Phase I. In Phase III, the mean age is close to Phase II with 46.76 ± 13.68 (average other clusters: 42.51 ± 12.55). The onset age also is close to its former level 35.69 ± 15.18 (average other clusters: 35.19 ± 13.90). And so does the Mini-TQ sum score at 13.93 ± 5.86 (average other clusters: 14.21 ± 5.85). So, in Phase II the adherent group has a slightly higher age and age on onset, but a similar Mini-TQ sum score and the non-adherent group has in average a lower age and age on onset and also a similar Mini-TQ sum score. In Phase III, the values did not change significantly in either group. The next focus is on the noise group, i.e., users who are not assigned to a cluster.

Comparing some variables, we find that the “noise” instances are younger by comparison, with a mean age of 39 years (STD: ±12.04 years) compared to the average of the other clusters, with a mean age of 44.60 years (mean STD: ±12.86 years). This is also true for the age of onset with 23.45 ± 19.00 years to an average of 36.23 ± 14.21 years. The Mini-TQ sum score is also below the average of the other clusters with 8.19 ± 8.16 to an average of 14.05 ± 5.74. Compared to the distribution in the whole dataset (m: 60.01%/f: 32.36%/u: 1.64%), the noise instances also show a different gender distribution: (m: 31.77%/f: 49.47%/u: 18.75%). In the Phase II adherence group, the mean age is higher at 41.13 ± 13.10, and the age at onset is also higher at 26.11 ± 20.34. The Mini-TQ sum score is more stable (8.86 ± 8.79). The gender distribution is m: 65.04%/f: 33.06%/u: 1.89% for all and for the noise group it is: m: 29.00%/f: 46.00%/u: 25.00%. In Phase III, the mean age decreases to 40.62 ± 13.29. The entry age also drops to 24.93 ± 21.03. And so does the Mini-TQ sum score at 8.27 ± 8.76. The gender distribution is m: 65.37%/f: 32.68%/u: 1.95% for all and for the noise group it is: m: 31.70%/f: 41.46%/u: 26.83%. In the non-adherence Phase II group, mean age is lower at 38.66 ± 10.94, also age at onset at 20.89 ± 17.38 and Mini-TQ sum score at 7.45 ± 7.40. The gender distribution is m: 67.44%/f: 31.28%/u:1.28% for all and for the noise group it is: m: 34.78%/ f: 53.26%/u: 11.96%. In Phase III, mean age increases again to 39.36 ± 11.25, age at onset to 22.46 ± 17.57, and Mini-TQ sum score to 8.13 ± 7.72. The gender distribution is m: 66.53%/f: 32.09%/u: 1.38% for all and for the noise group it is: m: 31.81%/f: 55.45%/u: 12.72%.

In the noise group, the proportion of females is significantly higher than in the overall group in Phase I. In Phase II & III, this dominance is maintained, but there are no strong deviations between the two adherence groups. The adherence group is minimally older and has a minimally higher age of onset. In addition, the Mini-TQ sum score tends to be slightly higher than in the non-adherence group, especially in Phase II.

4.4. Tuning the gap size for prediction

Our dataset contains many sequences with small gaps, but some sequences have gaps of more than 2000 days. We filtered out sequences with a longer gap than one month (30 days). This reduced the dataset size considerably, since only 10 users had gaps of 15 days, and larger gaps were even more rare. Hence, we limited the maximum gap size to 30 and invoked the three binning strategies described in Section 3.7, to build 5 groups of increasing gap size (number of days). The results are on Table 2. There, we sorted the 5 groups by size, with the group of the smallest gaps coming first (c.f. leftmost column). As can be seen in the table, the frequency-based strategy has built only three groups, the 3rd of which contains gaps of very different sizes, from 4 to 30 days. The other two strategies placed in each group sequences of similar gap sizes, whereupon the third (rightmost) strategy achieved a somehow smoother distribution of sequences among groups.

Applying 1-NN algorithm to the sequences in order to predict the gaps (based on Fischer–Jenks binning method) led to a mean accuracy of 61% for stratified 10-fold CV. The minority classes E & D performed worse as well as label C. In each fold, the Precision and Recall values are around 15% for label B and 77% for label A, which corresponds to their class distribution (c.f. Table 3 for the detailed evaluation metrics).

The 1-NN classifier with DTW was chosen because it is the gold standard in performance according to [9].

4.5. Discussion

Only Dynp identified the stages of attrition, according to Eysenbach’s [1] description. The L-shape of the curve, is due to the nature of the app. While Eysenbach and Hochheimer et al. [4] work with trial or survey data, the users here are in a voluntary exploration situation. The app is only used over a longer period of time if the relationship fits [1]. According to [4] “sensitive user-specified thresholds” are able to correctly identify dropout phases. We could demonstrate an approach with Change point Detection methods that requires less manual effort. This allows the identification of phases with different dropout rates to be integrated into a more autonomous workflow.

The prediction of Phase III from I+II outperforms the prediction of Phase II from I, Since this prediction is close to the class distribution. However, the result is explained by the available input. While the first prediction can only use a single value per user, the second prediction can use 4. The problems of the algorithm, although it can be considered ‘state-of-the-art’, can be attributed to the large number of users, with subjective manifestations of tinnitus. Moreover, the variable has a very similar value (Mean: 0.357, SDT: 0.27), which further complicates its separation. By changing the setting, e.g. in a clinical study, the results

Table 2
Gap bins derived by each of the binning strategies on a total of 3749 sequences: strategy of equisized intervals (left column), frequency-based strategy (middle column) and the Fischer–Jenks algorithm that builds natural groups (rightmost column): each entry contains an interval size and the number of intervals of this size, as found by the algorithm. The leftmost column contains the group ID.

GroupID	Equisized intervals		Frequency-based		Fischer–Jenks	
	Interval	Sequences	Interval	Sequences	Interval	Sequences
A: smallest gaps	(1.999–7.6]	3543	(1.9–3.0]	2903	(1.999–3.0]	2903
B: small gaps	(7.6–13.2]	138	(3.0–4.0]	307	(3.0–6.0]	580
C: larger gaps	(13.2–18.8]	40	(4.0–30.0]	539	(6.0–11.0]	61
D: large gaps	(18.8–24.4]	15			(11.0–18.0]	77
E: very large gaps	(24.4–30.0]	13			(18.0–30.0]	28

Table 3
This table shows the precision, recall, and F1-score as evaluation metrics for each class and each fold of the 10-fold stratified cross-validation as well as the averages for the 1-NN classifier (Pre—Precision, Rec—Recall, F1—F1-score, Sup—Support, Avg—Average & Acc—Accuracy)

	Class A				Class B				Class C				Class D				Class E				
Fold	Pre	Rec	F1	Sup	Pre	Rec	F1	Sup	Pre	Rec	F1	Sup	Pre	Rec	F1	Sup	Pre	Rec	F1	Sup	Acc
1	0.79	0.80	0.80	291	0.21	0.22	0.22	58	0.08	0.06	0.07	16	0.00	0.00	0.00	7	0.00	0.00	0.00	3	0.66
2	0.78	0.81	0.80	291	0.22	0.19	0.20	58	0.00	0.00	0.00	16	0.00	0.00	0.00	7	0.00	0.00	0.00	3	0.66
3	0.79	0.79	0.79	291	0.16	0.17	0.17	58	0.00	0.00	0.00	16	0.00	0.00	0.00	8	0.00	0.00	0.00	2	0.64
4	0.77	0.73	0.75	290	0.17	0.22	0.19	58	0.00	0.00	0.00	17	0.00	0.00	0.00	8	0.00	0.00	0.00	2	0.60
5	0.77	0.74	0.75	290	0.14	0.17	0.16	58	0.14	0.12	0.13	16	0.00	0.00	0.00	8	0.00	0.00	0.00	3	0.60
6	0.77	0.72	0.75	290	0.13	0.16	0.14	58	0.00	0.00	0.00	16	0.08	0.12	0.10	8	0.00	0.00	0.00	3	0.59
7	0.79	0.73	0.76	290	0.17	0.19	0.18	58	0.00	0.00	0.00	16	0.08	0.12	0.10	8	0.00	0.00	0.00	3	0.60
8	0.78	0.74	0.76	290	0.18	0.21	0.19	58	0.04	0.06	0.05	16	0.08	0.12	0.10	8	0.00	0.00	0.00	3	0.61
9	0.80	0.77	0.79	290	0.20	0.22	0.21	58	0.10	0.12	0.11	16	0.00	0.00	0.00	8	0.00	0.00	0.00	3	0.63
10	0.77	0.71	0.74	290	0.10	0.12	0.11	58	0.00	0.00	0.00	16	0.00	0.00	0.00	7	0.00	0.00	0.00	3	0.57
Avg	0.78	0.75	0.77	290.30	0.17	0.19	0.18	58.00	0.04	0.04	0.04	16.10	0.02	0.04	0.03	7.70	0.00	0.00	0.00	2.80	**0.62**

might be better, because the phases might be more sigmoid-shaped. Other studies may refer to direct prediction of adherence overall. For example, in [12], adherence to treatment is predicted with a minimum input of 90 days, further increasing in 30-day blocks [12]. Or in the work of [13], fitness program adherence is predicted. But again with an input of 90 days and a non-adherence definition of more than one month without training activity [13]. So the presented method has to cope with much less input, at least in the example chosen here. In addition, it offers the possibility to classify users earlier according to their condition, and consequently, it would be possible to react accordingly. As an example, long-term users may have different demands on an app than new users who first have to be won over for an app or program.

For the breakdown into clusters, it can be stated that the HDBSCAN algorithm was able to find stable clusters. In addition, it could be shown that there are different tendencies in the adherence behavior in the clusters and that the differences in the composition of the clusters are recognizable. Thus, it seems possible to predict the development of users by cluster membership. The results point in a positive direction and the approach should be pursued. In addition, the approach of using the Jaccard distance seems to be helpful in case of high non-adherence. The manual evaluation, on the other hand, is laborious and a more appropriate measure for cluster evaluation before analysis also seems desirable. There is potential for improvement in this direction. In summary, the clustering step adds value for the users of the method by identifying groups of users that, similar to the approach in the phase model, provide early indications of possible behavior of new users that can then be acted upon in terms of the app providers. Clustering is also applied in [13] to improve workflow results. However, their work uses K-Means, which requires the number of clusters to be defined in advance and also does not take noise in the data into account. Both of these make the application unsuitable in the present methodology. In addition, they use clustering to better train their regression models and not to identify user groups by their characteristics and to examine their adherence behavior over time in order to gain information from them independently.

Learning the gap labels by binning suffers from the skew in the gap size. Nevertheless, the Fischer–Jenks algorithm performed better than the other ones in creating reasonable groups. An adaptation of the number of labels or the time frame of gaps might also influence the result, but were out of scope for this exploratory paper and is future work.

The prediction of the gaps by 1-NN also indicates the aforementioned problems with the complex data. And points in a direction of methods that adjusts not only to the matching of the sequences, but also to the different subgroups of the user in order to achieve better predictive power. The results from the clustering step suggest this, as does the embedding in the workflow by [13]. Other work also seems to have problems with predicting adherence. For example, the authors of [12] have achieved a sensitivity of 0.81 for a dropout event after 90 days, but a specificity of 0.65 and this in a 2-class problem and much more input data [12]. The results of the evaluation require justification. This is due to the skewed distribution of the learned classes in the test data on the one hand and to the nature of the data itself on the other hand, since tinnitus is already by definition a very heterogeneous disorder and the patients are very different from each other. Under these conditions, even a different algorithm would perform below expectations.

5. Conclusions

In this paper, we proposed a method to investigate a mHealth dataset for varying phases of attrition according to Eysenbach and predicted if a user might reach the next phase from the current status. We used the fragmentation of the time series to predict users' pauses with the sequences, after determining them by binning.

We have found that the phases of attrition are best detected by the Dynp algorithm using change point detection. We have also shown that they can be predicted by XGBoost for many users, even under challenging data. The Fischer–Jenks algorithm excelled in detecting gap labels. In addition, the results of the predictions indicate that good alignment of the sequences is not sufficient to make good predictions on this data. And particularly useful for medical experts, we were able to show that clustering in combination with the phases with different dropout rates can identify groups of users who exhibit specific adherence patterns.

Limitation of the approach is the still low predictive power in view of the very heterogeneous users and the high fluctuation of users. The approach should still be tested on a dataset that offers more stable

conditions of use of the mHealth app, such as a clinical trial, a survey, or an experiment.

Future studies are the evaluation on datasets with other patterns, in-depth adjustments of the parameters of the individual tools and possibly testing new elements that harmonize better with the properties of the data.

Our approach points in a good direction. Open questions remain, however, regarding the quality of prediction in the domain of self-monitoring mHealth data from volunteers, an issue that warrants further investigation.

Ethical approval

#15-101-0204. "All users read and approved the informed consent before participating in the study. The study was carried out in accordance with relevant guidelines and regulation". (Ethics Committee of the University Clinic of Regensburg.)

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement number 848261.

References

[1] Eysenbach G. The law of attrition. J Med Internet Res 2005;7(1):e402.
[2] World Health Organization and others. Adherence to long-term therapies: Evidence for action. World Health Organization; 2003.
[3] Hochheimer CJ, Sabo RT, Krist AH, Day T, Cyrus J, Woolf SH. Methods for evaluating respondent attrition in web-based surveys. J Med Internet Res 2016;18(11):e301.
[4] Hochheimer CJ, Sabo RT, Perera RA, Mukhopadhyay N, Krist AH. Identifying attrition phases in survey data: applicability and assessment study. J Med Internet Res 2019;21(8):e12811.
[5] Cismondi F, Fialho AS, Vieira SM, Reti SR, Sousa JM, Finkelstein SN. Missing data in medical databases: Impute, delete or classify? Artif Intell Med 2013;58(1):63–72.
[6] Schleicher M, Unnikrishnan V, Neff P, Simoes J, Probst T, Pryss R, et al. Understanding adherence to the recording of ecological momentary assessments in the example of tinnitus monitoring. Sci Rep 2020;10(1):1–13.
[7] Williams-Kerver GA, Schaefer LM, Hazzard VM, Cao L, Engel SG, Peterson CB, et al. Baseline and momentary predictors of ecological momentary assessment adherence in a sample of adults with binge-eating disorder. Eat Behav 2021;41:101509.
[8] Schleicher M, Pryss R, Schobel J, Schlee W, Spiliopoulou M. Expect the gap: A recommender approach to estimate the absenteeism of self-monitoring mhealth app users. In: 2022 IEEE 9th international conference on data science and advanced analytics. 2022, p. 1–10.
[9] Bagnall A, Lines J, Bostrom A, Large J, Keogh E. The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances. Data Min Knowl Discov 2017;31.
[10] Schleicher M, Hamacher S, Naujoks M, Günther K, Schmidt T, Pryss R, et al. Prediction of declining engagement to self-monitoring apps on the example of tinnitus mhealth data. In: 2022 IEEE 35th international symposium on computer-based medical systems. IEEE; 2022, p. 228–33.
[11] Puga C, Schleicher M, Niemann U, Unnikrishnan V, Boecking B, Brueggemann P, et al. Juxtaposing medical centers using different questionnaires through score predictors. Front Neurosci 2022;16:193.
[12] Gottlieb A, Yatsco A, Bakos-Block C, Langabeer JR, Champagne-Langabeer T. Machine learning for predicting risk of early dropout in a recovery program for opioid use disorder. Healthcare 2022;10(2):223.
[13] Jossa-Bastidas O, Zahia S, Fuente-Vidal A, Sánchez Férez N, Roda Noguera O, Montane J, et al. Predicting physical exercise adherence in fitness apps using a deep learning approach. Int J Environ Res Public Health 2021;18(20):10769.
[14] Schlee W, Pryss RC, Probst T, Schobel J, Bachmeier A, Reichert M, et al. Measuring the moment-to-moment variability of tinnitus: the trackyourtinnitus smart phone app. Front Aging Neurosci 2016;8:294.
[15] Cederroth CR, Gallus S, Hall DA, Kleinjung T, Langguth B, Maruotti A, et al. Towards an understanding of tinnitus heterogeneity. Front Aging Neurosci 2019;11:53.
[16] Killick R, Fearnhead P, Eckley IA. Optimal detection of changepoints with a linear computational cost. J Amer Statist Assoc 2012;107(500):1590–8.
[17] Keogh EJ, Chu S, Hart D, Pazzani MJ. An online algorithm for segmenting time series. In: Proceedings of the 2001 IEEE international conference on data mining. 2001, p. 289–96.
[18] Fryzlewicz P. Unbalanced haar technique for nonparametric function estimation. J Amer Statist Assoc 2007;102(480):1318–27.
[19] Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM; 2016, p. 785–94.
[20] Hiller W, Goebel G. Rapid assessment of tinnitus-related psychological distress using the mini-TQ. Int J Audiol 2004;43(10):600–4.
[21] Langguth B, Goodey R, Azevedo A, Bjorne A, Cacace A, Crocetti A, et al. Consensus for tinnitus patient assessment and treatment outcome measurement: Tinnitus research initiative meeting, Regensburg, july 2006. Prog Brain Res 2007;166:525–36.
[22] Hallam R. Manual of the tinnitus questionnaire (TQ). London: Psychological Corporation; 1996.
[23] Hiller W, Goebel G. A psychometric study of complaints in chronic tinnitus. J Psychosom Res 1992;36(4):337–48.
[24] Kojima T, Kanzaki S, Oishi N, Ogawa K. Clinical characteristics of patients with tinnitus evaluated with the tinnitus sample case history questionnaire in Japan: A case series. PLoS One 2017;12(8):e0180609.
[25] McInnes L, Healy J, Astels S. HDBSCAN: Hierarchical density based clustering. J Open Source Softw 2017;2(11):205.
[26] Campello RJGB, Moulavi D, Sander J. Density-based clustering based on hierarchical density estimates. Lecture Notes in Artificial Intelligence 2013;7819:160–72.
[27] Jenks GF. The data model concept in statistical mapping. Int Yearb Cartogr 1967;7:186–90.