Adaptive data collection for intraindividual studies affected by adherence

Greta Monacelli¹,² | Lili Zhang¹,² | Winfried Schlee³ | Berthold Langguth³ | Tomás E. Ward¹,² | Thomas B. Murphy⁴, 5

¹School of Computing, Dublin City University, Dublin, Ireland
²Insight SFI Research Centre for Data Analytics, Dublin, Ireland
³Department of Psychiatry and Psychotherapy, University of Regensburg, Regensburg, Germany
⁴School of Mathematics and Statistics, University College Dublin, Dublin, Ireland

Abstract

Recently, the use of mobile technologies in ecological momentary assessments (EMAs) and interventions has made it easier to collect data suitable for intraindividual variability studies in the medical field. Nevertheless, especially when self-reports are used during the data collection process, there are difficulties in balancing data quality and the burden placed on the subject. In this paper, we address this problem for a specific EMA setting that aims to submit a demanding task to subjects at high/low values of a self-reported variable. We adopt a dynamic approach inspired by control chart methods and design optimization techniques to obtain an EMA triggering mechanism for data collection that considers both the individual variability of the self-reported variable and of the adherence. We test the algorithm in both a simulation setting and with real, large-scale data from a tinnitus longitudinal study. A Wilcoxon signed rank test shows that the algorithm tends to have both a higher $F_1$ score and utility than a random schedule and a rule-based algorithm with static thresholds, which are the current state-of-the-art approaches. In conclusion, the algorithm is proven effective in balancing data quality and the burden placed on the participants, especially in studies where data collection is impacted by adherence.

KEYWORDS

control charts, design optimization, ecological momentary assessments, intraindividual studies

1 | INTRODUCTION

In recent years, mobile technologies have made longitudinal data in the medical field more easily available and, at the same time, allowed to implement more complex and challenging experimental designs. Mobile applications or activity trackers can be used to collect multiple sensory data or self-reports for long periods of time. These new technologies have...
been effectively applied both in the behavioral field, through ecological momentary assessments (EMAs) and interventions (EMIs) (de Vries et al., 2021), and in clinical trials, through electronic patient-reported outcomes or ePRO systems. These outside-the-lab experiments allow for more complex research goals and designs. For example, it is possible to send notifications through a mobile application to the participants based on the recognition of a relevant context such as symptoms severity or physical location (de Vries et al., 2021; Hekler et al., 2018; Hulme et al., 2021). However, as the complexity of the experiments increases, the logistic challenges to correctly deliver them also increase and efficiency in data collection becomes necessary to reduce the costs of a lengthy experiment. The term costs may refer to different downsides of the experiment based on the context; examples are the energy consumption of the mobile application (Hekler et al., 2018) or the burden placed on the subjects during the data collection process.

In this context, intrindividual studies have gathered increasing attention (Liu et al., 2019; Russell et al., 2020). Intrindividual variability is an important aspect in medical research. For example, it has been used to detect neurological and psychological conditions (MacDonald et al., 2006). Moreover, it can negatively impact medical studies as differences between groups may become less evident due to high intrindividual variability, especially for small sample-sized studies (MacDonald et al., 2006). The type of data collected in the outside-the-lab experiments mentioned above can be used for intrindividual variability studies. Indeed, for each subject, multiple observations of the same variable are collected over time. Moreover, the more complex designs made available by mobile technologies can be also applied. This allows to optimize the data collection process and focus on the intrindividual differences of a variable when specific conditions are satisfied. In this work, we focus on an EMA that aims to submit a demanding task to subjects at high/low values of a self-reported variable using a mobile application. The demanding task is assumed to be burdensome or costly for either the participants or the researcher, so that triggering it for every interaction of the subject with the application is not feasible.

An example of such EMA setting for tinnitus patients is what motivated this work (Zhang et al., 2022). Tinnitus has great repercussions on the quality of life of chronic affected patients (Probst et al., 2016). Thus, evaluating the impact of this condition on cognitive processes and decision-making is of interest. Hence, the EMA study aims to collect both tinnitus self-reports and decision-making data from a group of tinnitus patients and compare their performance at an intrindividual level. Multi-armed bandit tasks are used to collect the decision-making data. These tasks have been used effectively in the emerging field of computational psychiatry to evaluate human decision-making and are considered potential new biomarkers for psychiatric and psychological conditions (Ahn et al., 2016). Nevertheless, they are relatively lengthy and the collection of such data can be burdensome on the subjects. Therefore, we seek to trigger these tasks only for high and low values of tinnitus severity. This decreases the number of multi-armed bandit tasks that needs to be triggered during the study for each subject while still allowing an intrindividual comparison at meaningful values of tinnitus severity. Notice that the experiment described here can be easily generalized: the multi-armed bandit task can be replaced by a generic additional task that has great cost either for the researcher (expensive equipment) or for the subject (lengthy or invasive); similarly, researchers can choose the self-reported variable based on their research question.

There are two possible immediate, but sometime ineffective, solutions to this sampling problem. The first option is a random algorithm that randomly selects when to trigger the additional tasks. The second option is a static algorithm that triggers the additional tasks for values of the self-reported variable above or below predetermined thresholds. Several studies in the behavioral field have implemented either the first or the second option in similar contexts to the one tackled in this paper (de Vries et al., 2021). These sampling strategies are relatively easy to implement, but they do not consider three major problems which, instead, we ought to tackle. First, the expected value and variance of the self-reported variable may vary across subjects. For example, as we know from prior studies, tinnitus patients present a diversity of self-reported symptom ranges (Probst et al., 2016). Second, it is necessary to balance the necessity of the researcher to collect enough data with the risk of overburdening the subjects with a too intensive data collection process. Third, we need to consider low adherence to the study. This is important especially when self-reports are considered as they place additional burden on the subjects and can lead to both incomplete data and early drop-out from the experiment. Hence, the former triggering strategies may not trigger at all for some patients and excessively trigger for others. Moreover, there could be cases in which only additional tasks for high values of the self-reported variable are triggered and none for low values (or vice versa). In these cases, we expect to require much longer experiments to increase the likelihood of capturing data at the extremes targeted. This reason necessitated the development of the adaptive triggering algorithm described here.

In this work, we develop an algorithm that tackles the three problems mentioned above. The algorithm is heavily inspired by control chart approaches and adapts to the individual subject based on their past history of self-reports (Fricker et al., 2008). Moreover, it uses a design optimization method to change the definition of high and low values for the self-reported variable in an adaptive way, based on the expected adherence of the subject and on the balance between the burden and the efficacy of the experiment. Thus, the algorithm is designed to collect data as soon as possible for nonad-
herent subjects and, in contrast, to burden the adherent ones only when strictly necessary. The algorithm outperforms, according to our metrics, the state-of-the-art approaches—random schedule and the static algorithm—for both simulated and real tinnitus data; moreover, our results suggest that the algorithm is particularly effective when the adherence of the subjects is highly variable—a very common situation in EMA studies that rely on self-reports.

1.1 Previous work

Several adaptive methods have been developed to improve data collection in EMA and EMI, but none focuses on the delivery of an additional and demanding task for a predetermined number of times at extreme self-reported values or on the explicit use of the adherence history to improve the algorithm. For example, Mohan (2021) and Hekler et al. (2018) increase the efficacy of EMI through adaptive rule-based algorithms. Hekler et al. (2018) suggest a statistical approach to decrease the energy consumption of the mobile application used in the study while maintaining the effective monitoring of a specific heart condition. Additionally, Thomas and Bond (2015) and Hulme et al. (2021) have developed solutions that deliver just-in-time notifications to the participants through approaches based, respectively, on predetermined rules and hidden Monte Carlo methods.

Previous work from two different fields has been combined to obtain the results in this paper: anomaly—or outlier—detection and design optimization. A similar approach has been used by Koizumi et al. (2012) to search for an optimal significance level in an anomaly detection problem. First, we looked at the anomaly/outlier detection literature to identify extremes values for the self-reported variable. Identifying unusual patterns in the data concerns many different applications and there is a vast literature both in the computational field—anomaly detection (Chandola et al., 2009)—and in the statistical field—outliers detection (Zimek et al., 2012). Second, we used a design optimization perspective to obtain a more balanced definition of extreme values that considers the adherence of the subjects, that is, the sample size of the data they provide during the experiment. Design optimization methods have been applied in various contexts to balance data quality and experimental costs: examples include clinical trials for drug testing (Sylvester, 1988) and experiments in computational psychiatry (Cavagnaro et al., 2010).

2 METHOD

Let us consider an \(N_d\) days long experiment. During this time, a participant of the study is asked to self-report a quantity of interest \(N_h\) times per day through notifications pushed to their mobile phone. The time at which these notification are sent can be either prefixed or randomized. At the end of the experiment, the participant will have received a total of \(N = N_d N_h\) notifications. The data is collected as a time series \(x = \{x_t\}_{t=1}^N\), where \(x_t \in [0, 1] \cup \{\text{NaN}\}\), \(t \in \{1, \ldots, N\}\) is the discrete time, and NaN denotes missing values. We wish to trigger an additional task for high and low values of the quantity of interest \(x_t\).

2.1 Control chart

Let us first consider the simplified case of a completely adherent participant, so that the time series \(x\) does not contain any missing values (NaN). Assume that \(x\) is an i.i.d. sample from a random variable \(X \sim \text{Beta}(\delta, \xi)\), that is, a Beta distribution of parameters \(\delta\) and \(\xi\). Let the triggering starting point \(S \in \{2, \ldots, N\}\) be the first time point after which the additional task can be triggered. For every \(s \in \{S - 1, \ldots, N\}\), consider the unbiased estimators of the expected value and the variance of the subsample \(x_{1:s} = \{x_t\}_{t=1}^s\), that is,

\[
\hat{\mu}_s = \frac{1}{s} \sum_{r=1}^{s} x_r \quad \text{and} \quad \hat{\sigma}^2_s = \frac{1}{(s - 1)} \sum_{r=1}^{s} (x_r - \hat{\mu}_s)^2.
\]

Assume that, for all \(s \in \{S - 1, \ldots, N\}\),

\[
0 < \hat{\sigma}^2_s < \hat{\mu}_s (1 - \hat{\mu}_s).
\]
Then, the estimates of the parameters $\delta$ and $\xi$ based on the subsample $x_{1:s}$ obtained through the method of moments are

$$\hat{\delta}_s = \hat{\mu}_s \nu_s \quad \text{and} \quad \hat{\xi}_s = (1 - \hat{\mu}_s) \nu_s,$$

(2)

where $\nu_s = \hat{\mu}_s (1 - \hat{\mu}_s) / \hat{\sigma}_s^2 - 1$. The method of moments was preferred to the maximum likelihood estimation due to computational reasons; indeed, there is no closed-form solution for the latter option.

Let $\alpha \in [0, 1]$ be the significance level. Then, we say that the observation $x_t$ is extreme if

$$x_t < z_{\alpha/2}(\hat{\delta}_{t-1}, \hat{\xi}_{t-1}) \quad \text{or} \quad x_t > z_{1-\alpha/2}(\hat{\delta}_{t-1}, \hat{\xi}_{t-1}),$$

(3)

where $t \geq S$ and $z_{\alpha}(\delta, \xi)$ is the $\alpha$-quantile of the Beta distribution with parameters $\delta$ and $\xi$. If Condition (3) is satisfied at time $t$, then the additional task is triggered.

In the anomaly detection and the outlier detection literature, the control chart approach is usually used to identify observations that are unlikely under a probability model, that is, outliers. Nevertheless, it can be applied in the context of this paper by assuming that no outliers or anomalies are present in the data. Indeed, under this assumption, the values identified through Equation (3) are considered valid even though unlikely. Hence, they can be considered extremes because of their position “close” to the boundaries of $[0,1]$, outside of the interval $[z_{\alpha/2}(\hat{\delta}_{t-1}, \hat{\xi}_{t-1}), z_{1-\alpha/2}(\hat{\delta}_{t-1}, \hat{\xi}_{t-1})]$.

### 2.2 Selection of the significance level and the triggering starting point

In Section 2.1, the criteria to select the significance level $\alpha$ and the triggering starting point $S$ are not specified. In the outliers and anomaly detection literature, both of these design variables are usually chosen by the researcher based on domain knowledge or previous studies (Chandola et al., 2009). In this section, we investigate this choice in the formal framework of design optimization in order to aid the researcher in the final selection (Chaloner & Verdinelli, 1995) and following a similar approach to Koizumi et al., 2012.

First, we define a utility function that represents the goal of the researcher. We assume that, in order to maintain a balance between the burden placed on the subjects and the necessity of the researcher to collect enough data, the primary goal of the study is to trigger the additional task on average $v \in \{1, \ldots, N\}$ times for each participant.

To define the utility, let us consider the time series $w = \{w_t\}_{t=S}^{N}$ defined as

$$w_t = \begin{cases} 1 & \text{if the control chart triggers the additional task at time } t \\ 0 & \text{otherwise.} \end{cases}$$

Assume that $w$ is an i.i.d. sample from a Bernoulli random variable $W \sim \text{Ber}(\alpha)$. Let the random variable $V$ represent the total number of additional tasks triggered during the experiment. Then, $V \sim \text{Bin}(\alpha, N - S + 1)$ follows a Binomial distribution. Hence, the expected value of $V$ is equal to $E(V) = (N - S + 1)\alpha$. We choose as utility function $U_1 : \{2, \ldots, N\} \times [0, 1] \to \mathbb{R}$ defined as

$$U_1(S, \alpha) = -(E(V) - v)^2.$$

Thus, $U_1(S, \alpha)$ associates higher value to experimental designs $(S, \alpha)$ for which the expected number of triggers $E(V)$ is as close as possible to the desired one $v$.

Let $D = \{(S, \alpha) \in \{2, \ldots, N\} \times [0, 1]\}$, then we aim to find an optimal design $(S^*, \alpha^*)$ such that

$$(S^*, \alpha^*) \in \arg \max_{(S, \alpha) \in D} U_1(S, \alpha).$$

(4)

The design optimization problem in Equation (4) has as solutions

$$A^* = \{(S^*, \alpha^*) \in D | \alpha^* = \max \left(0, \min \left(1, \frac{v}{(N - S^* + 1)}\right)\right)\}.$$

(5)

Therefore, experimental designs chosen in the set $A^*$ are all optimal for collecting, on average, $v$ additional tasks.

Smaller values of the significance level $\alpha$ should be preferred because they result in more significant intraindividual comparisons at the end of the experiment. If $\alpha^*$ is chosen to be small, then, by the constraint in Equation (5), the starting
Algorithm 1

1: Set $N \in \mathbb{N}^+$ and $S^*, v \in [2, \ldots, N]$. Set $\alpha = \max(0, \min(1, \frac{v}{(N - S^* + 1)}))$.
2: for $t = 1 \ldots, N$ do
3: collect $x_t$
4: if $t \geq S^* \text{ and } x_t \neq NaN$ then
5: compute $\delta_{t-1}$ and $\xi_{t-1}$ as defined in Equation (2)
6: if Condition (3) holds then
7: trigger the additional task
8: end if
9: end if
10: end for

point $S^*$ is also small. However, while smaller values of $S^*$ can be useful in case of low adherence, they will also likely result in worse data quality because control charts rely on past observations to improve over time.

We suggest to first fix $S^*$ based on the domain knowledge of adherence in the field and, in any case, to not choose too small a value as it can negatively impact the estimation of the parameters $\delta_s$ and $\xi_s$. Once $S^*$ has been fixed, Equation (5) gives the optimal significance level $\alpha^*$, that is,

$$
\alpha^* = \begin{cases} 
0 & \text{if } N - S^* + 1 < 0 \\
1 & \text{if } 0 \leq N - S^* + 1 \leq v \\
\frac{v}{(N - S^* + 1)} & \text{otherwise.}
\end{cases}
$$

Combining Equation (5) and the control chart in Section 2.1, we obtain Algorithm 1. Notice that, since the thresholds in the control chart are symmetric with respect to the probability, Algorithm 1 triggers on average $v/2$ additional tasks for both high and low values of the quantity of interest.

2.3 Adherence

The selection of the significance level in Section 2.2 relies on complete adherence from the subjects to the EMA. Indeed, the definition of $\alpha^*$ in Equation (6) assumes prior knowledge of the final number of samples $N$. Equivalently, missing data are not considered. Nevertheless, this assumption is almost never satisfied in clinical longitudinal studies that rely on self-reports (de Vries et al., 2021).

Assume that the data set has missing data, then the final total number of samples for the quantity of interest is $N' < N$. In this case, Algorithm 1 triggers on average a lower number of additional tasks than $v$. Indeed, if $N' > v + S - 1$, using Equation (6), we obtain that

$$
\alpha^* = \frac{v}{N' - S + 1} < \frac{v}{N - S + 1}.
$$

Thus, the optimal significance level for the missing data scenario should be larger than the one selected assuming complete adherence to trigger a similar number of additional tasks.

In order to tackle this problem, we can replace $N$ in Algorithm 1 with an estimator of $N'$. If data from a pilot study is available, it is possible to use a simple estimate of the total final number of samples. For example, let $B$ be the set of all the subjects in the pilot study and $N'_j$, $j = 1, \ldots, |B|$ be the total number of samples collected for each subject $j$. Then, $N'$ can be estimated by the average number of samples per subject, that is, $\overline{N'} = \frac{1}{|B|} \sum_{j \in B} N'_j$. In this paper, we used the tinnitus data set to obtain the estimate $\overline{N'}$ and, in the following, we will always assume that Algorithm 1 uses it.

Nevertheless, pilot studies are not always available. In those cases, it may be difficult to predict accurately the average adherence of the subjects. Moreover, even if the data is available, small sample size, external factors, interventions, and individual differences may decrease the accuracy of the algorithm.
Algorithm 2

1: Set $N \in \mathbb{N}^+$ and $S^*, v, R \in \{2, \ldots, N \}$.
2: for $t = 1 \ldots, N$ do
3:   collect $x_t$
4:   if $t \geq S^*$ and $x_t \neq NaN$ then
5:     compute $\hat{\delta}_{t-1}$ and $\hat{\xi}_{t-1}$ as defined in Equation (2)
6:     compute $\hat{N}'(t)$ and $\alpha^*(t)$ as defined in Equations (9) and (10)
7:   if Condition (3) holds and $\sum_{i=1}^{t-1} w_i \leq R$ then
8:     trigger the additional task
9:   end if
10: end if
11: end for

Another possible approach is to infer a simple adaptive estimate of $N'$ at each time point by considering only the history of adherence of the individual and to replace it in Equation (6). The advantage of this method is that the estimation of $N'$ does not require prior external data.

Let

$$a_t = \begin{cases} 0 & \text{if } x_t = \text{NaN} \\ 1 & \text{if } x_t \neq \text{NaN}, \end{cases}$$

so that the time series $a = \{a_t\}_{t=1}^{N}$ represents the adherence of a subject.

Assume that a participant decides to report the quantity of interest randomly and let the r.v. $A \sim Ber(\chi)$ represent the participant behavior. A success represents an interaction with the app, whereas a loss represents a missing data. Then, the number of successes $N'$, that is, the total number of samples we collect from that subject at the end of the experiment, follows a Binomial distribution of parameters $\chi$ and $N$. In particular,

$$E(N') = \chi N. \tag{7}$$

We can estimate the probability of success $\chi$ adaptively through the method of moments, thus obtaining

$$\hat{\chi}(s) = \frac{\sum_{i=1}^{s} a_i}{s}, \quad S^* - 1 \leq s \leq N'. \tag{8}$$

Then, replacing $\chi$ in Equation (7) provides the following estimator for the number of final samples $N'$, that is,

$$\hat{N}'(t) = \hat{\chi}(t - 1)N. \tag{9}$$

Thus, we obtain the following adaptive definition for the significance level, where $\alpha$ is now a function of the time $t$:

$$\alpha^*(t) = \begin{cases} 0 & \text{if } \hat{N}'(t) - S^* + 1 < 0 \\ 1 & \text{if } 0 \leq \hat{N}'(t) - S^* + 1 \leq v \\ \frac{v}{(N'(t) - S^* + 1)} & \text{otherwise.} \end{cases} \tag{10}$$

In conclusion, we obtain Algorithm 2.

Notice that the definition of extreme values in Algorithm 2 depends on the adherence of the participant. More adherent participants are expected to have lower values of the adaptive significance level $\alpha^*(t)$ than less adherent ones. Thus, the additional tasks will be triggered for more significantly unlikely/extreme values of $x_t$ if the participant is adherent. In contrast, the algorithm tends to collect data even if it is not highly significant/extreme from a statistical viewpoint for nonadherent subjects. This characteristic can be both an advantage and a disadvantage. It is an advantage as Algorithm 2 adapts to the adherence of the subjects as well as to the history of the quantity of interest. On the other hand, it is a
FIGURE 1  Graphical representation of the random algorithm, the static algorithm, Algorithms 1 and 2. The blue line represents the time series of the self-reported tinnitus severity for a synthetic patient. We set $S^* = 6$. Dots colored in white represent not-triggered tasks, while those colored in blue represent triggered ones. The blue area is the region of nonextreme values. In the case of the static algorithm, this consists in all observations between 0.15 and 0.85. On the other hand, for Algorithms 1 and 2, it represents the confidence intervals defined adaptively by the algorithms during the data collection process. The yellow area represents the ground truth. Its thresholds are computed using Equation (3) and selecting the optimal significance level $\alpha^*$ based on the true adherence of the subjects—known at posteriori at the end of the experiment. In particular, $\alpha^*$ is chosen replacing in Equation (6) the number of notifications $N$ with the number of samples collected for the individual subject $N'$. Therefore, triggered tasks outside this area are defined as true positives, and triggers inside are false positive. Nontriggered tasks inside the area are true negatives, and nontriggered tasks outside the area are false negatives. Notice that the confidence intervals of Algorithm 2 are smaller than those of Algorithm 1.

3  RESULTS

In this section, we compare a random and a static schedule with Algorithms 1 and 2 on both a simulation setting and on real data from a tinnitus longitudinal study. The random schedule randomly selects 10 interactions of the user with the app to trigger the additional task. If the user does not respond to the self-report at the selected times, the additional task trigger shifts to the next interaction of the user with the application until a separate trigger is set. The static algorithm is a rule-based approach with prefixed thresholds at 0.15 and 0.85. All results in this paper were produced using R (Team, 2021). The figures were obtained using the plotly R package (Sievert, 2020).

In both the simulation and the real data analysis, $v = 4$ and $N = 180$ are fixed. Similarly, the triggering starting point is set to $S^* = 6$. The latter decision was taken by considering the adherence in the real data and trying to balance two contrasting necessities. On the one hand, it is important to start collecting the additional task early as the adherence of the subjects decreases exponentially over time, see Figure 2. On the other hand, both Algorithms 1 and 2 depend on past observations and are not reliable without some prior data. In particular, the method of moments used for the parameter estimation may encounter computational errors if $S^*$ is too small. For more details on the choice of the starting point, we refer to the Supporting Information.

3.1  Simulated data

We simulated two data sets. In both cases, the data were generated from the statistical model used to develop Algorithm 2 so that the adherence follows a Binomial distribution and the quantity of interest follows a Beta distribution. The Beta distribution parameters for the simulated participants are sampled from a uniform distribution in the interval $[0.5, 10]$. However, in the first simulated data, the adherence is considered alike for all subjects. In particular, the total number
FIGURE 2  Comparison between the total number of observations per subject in the real data and the simulated data. Notice that some participants in the real data set have a higher number of observations than expected as they interacted with the application more than six times per day. In those cases, all samples are considered to estimate the parameters of the Beta distribution and an additional task can be triggered at those time points.

of samples for each subject is sampled from a Binomial distribution of parameters $N = 180$ and $\chi = 0.19$. On the other hand, in the second simulated data, more variability is added to the adherence by sampling for each subject a probability of success $\chi$ from a Beta$(1, 4)$ distribution $B(1, 4)$. As shown in Figure 2, this results in a data set more similar to the real data.

3.2  Real data

To investigate the performance of the algorithms in a real scenario, we consider a data set collected during a tinnitus-focused study where moment-to-moment tinnitus symptoms were reported using a mobile application called TrackYourTinnitus (Probst et al., 2016). Since intraindividual fluctuations may be an important characteristic feature of an individual’s tinnitus and it is of great importance to investigate the behavior changes while the individual is suffering from tinnitus, this data set is with significant practical value to test our triggering algorithms.

From the original data, we considered only the user_id, save, save_date, and question_2 columns. These columns collect, respectively, the users identification number, the time and/or date on which the user interacted with the TrackYourTinnitus application and the self-reported tinnitus severity. The tinnitus severity is represented by a value between 0 and 1. From this subset of the original data, we delete all rows that have at least a missing value and all duplicates. Occasionally, different tinnitus severity values were reported by the same user at the same identical time; in those cases, we only consider the first-listed of such observations. Since Algorithms 1 and 2 require a prefixed length of the experiment, we further restrict our attention to only the 30 days after the first interaction of each participant with the application. For most of the participants (77%, or $2130/2752$), there is no information loss as they interacted with the application only within this timeframe. Finally, as the starting point has been fixed to $S^\star = 6$, we do not consider users that interacted less than six times with the application. Notice that, after cleaning the data, only around 33% ($911/2752$) of the initial users remain.

3.3  Statistical analysis

To compare the performances of the four algorithms, both the distribution of the $F_1$ scores (Sokolova & Lapalme, 2009) and the estimated utility $u_1 = -\left(\sum_{i=1}^{N} w_i - v\right)^2$ were computed. The $F_1$ score measures the accuracy of the algorithms against the ground truth, which is defined based on Equations (3) and (6). This implies that the ground truth is tailored to the individual subject and varies across individuals, see Figure 1. On the other hand, the utility $u_1$ measures how effectively the algorithms achieve the goal set by the design optimization problem. The empirical cumulative distribution functions (eCDF) of these quantities for both the simulated data and the real data are shown in Figure 3. Moreover, we compare the eCDFs pairwise through the Wilcoxon signed rank test (Brunner et al., 2018) using the wilcoxon.test function in R. Notice that the computation of the $F_1$ score produced some NaN, which were omitted in the statistical test. The results are listed in Table 1.
We find that, for both the simulated and the real data, Algorithm 2 tends to have larger $F_1$ scores and utility values $u_1$ than all the other options. In particular, Algorithm 2 performs as well or better than Algorithm 1 but, unlike the other option, does not rely on any prior knowledge of the population adherence. Moreover, in the comparison between Algorithms 2 and 1, we obtain smaller $p$-values in the second simulated case compared to the first one, see the last column of Table 1. This discrepancy may be caused by the greater variability of adherence in the second simulated data, see Figure 2. The $p$-values obtained for the real data are also consistent with this hypothesis. Indeed, the real data has higher variability than the first simulated data but, unlike the second simulated data, were collected in a nonideal context. This suggests that the additional complexity of Algorithm 2 compared to Algorithm 1 is more “useful” for studies where there is a high variability in the adherence of the subjects.

4 | DISCUSSION

In conclusion, we have presented a valid alternative to the current state-of-the-art sampling methods for adaptive EMA based on extreme self-reports. Algorithms 1 and 2 have been proven effective to collect data through an additional and burdensome task based on high or low values of a less intrusive self-report. Indeed, they have both outperformed a static and a random triggering approaches in a simulated scenario. The same results are confirmed in the real case scenario of tinnitus severity data. While Algorithm 1 is easier to implement than Algorithm 2, the latter does not rely on prior knowledge of the population adherence and achieves similar or better results. Moreover, especially for studies with high
variability in adherence, the analysis suggests that Algorithm 2 is slightly more effective than Algorithm 1 in increasing data quality—obtaining a desired number of additional tasks for a higher value of $u_1$—while preserving precision—similar $F_1$ scores. This is especially important as many fields, such as the tinnitus research area, rely on self-reports for which adherence is a challenging problem.

Algorithms 1 and 2 can be easily adapted to various contexts, but still rely on strong statistical assumptions that may not hold in real scenarios. It is possible to generalize these algorithms to other distributions: arguments in the paper remain valid, with adequate changes to the notation, if the quantity of interest $X$ follows an absolutely continuous distribution. Nevertheless, it is still necessary to have easy access to the quantile of the distribution in order to effectively deploy the algorithms. On the other hand, we assume that the distribution of the quantity of interest $X$ is known a-priori. In addition, the quantity of interest and the adherence of the subjects are assumed independent and identically distributed even though time is often an important factor in longitudinal studies. Despite the latter limitations, these algorithms can be easily adapted to the necessity of the researcher and we have demonstrated their effectiveness through simulations and large-scale real-world analysis.

4.1 Future work

There are many possible ways to further develop the arguments in this paper. Using statistical models that consider time for both the quantity of interest and the adherence, such as autoregressive models, could improve the performance of the algorithms in real case scenarios (Fricker et al., 2008). Moreover, more than one quantity of interest is often collected during a longitudinal study (Probst et al., 2016) and there may be an interest in developing a multivariate version of the algorithms (Chandola et al., 2009). In particular, including objective measures, such as measurements from wearable sensors, may increase their effectiveness (MacDonald et al., 2006). In this direction, hierarchical methods could inspire future work as they have been successfully applied in the field of anomaly detection (Sottas et al., 2011). Finally, the algorithms developed in this paper provide an adaptive sampling schedule for deployment of a secondary burden-heavy task contingent on data from a lighter, more frequent data collection task. Nevertheless, it does not address the sampling scheme for the lighter task. An adaptive sampling scheme that predicts the unlikely states and sends a notification to the participants only when needed could further increase the balance between data quality and the burden placed on the subjects (Hulme et al., 2021).

ACKNOWLEDGMENTS

This research is supported by Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289_P2, cofunded by the European Regional Development Fund.

Open access funding provided by IReL.

CONFLICT OF INTEREST

The authors have declared no conflict of interest.

DATA AVAILABILITY STATEMENT

The code, the simulated data, and supplementary material about the implementation of the algorithms are available from the corresponding author or on our lab GitHub repository (https://github.com/AI-for-Better-Living/adc-tinnitus). Due to privacy and ethical reasons, the data from the TrackYourTinnitus application is not publicly available. It is available on request (winfried.schlee@ieee.org).

OPEN RESEARCH BADGES

This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available in the Supporting Information section.

This article has earned an open data badge “Reproducible Research” for making publicly available the code necessary to reproduce the reported results. The results reported in this article were reproduced partially due to confidentiality issues.

ORCID

Greta Monacelli https://orcid.org/0000-0003-4677-4661
Thomas B. Murphy https://orcid.org/0000-0002-5668-7046
REFERENCES

Ahn, W. Y., Dai, J., Vassileva, J., Busemeyer, J. R., & Stout, J. C. (2016). Computational modeling for addiction medicine. In H. Ekhtiari and M. P. Paulus (Eds.), Progress in brain research (vol. 224, pp. 53–65). Elsevier. https://linkinghub.elsevier.com/retrieve/pii/S0079612315001387

Brunner, E., Bathke, A. C., & Konietschke, F. (2018). Rank and pseudo-rank procedures for independent observations in factorial designs: Using R and SAS, 1st ed. Springer Series in Statistics. Springer, 2018.

Cavagnaro, D. R., Myung, J. I., Pitt, M. A., & Kujala, J. V. (2010). Adaptive design optimization: A mutual information-based approach to model discrimination in cognitive science. *Neural Computation*, 22(4), 887–905. https://www.mitpressjournals.org/doi/abs/10.1162/neco.2009.02-09-95

Chaloner, K., & Verdinelli, I. (1995). Bayesian experimental design: A review. *Statistical Science*, 10(3), 273–304.

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys*, 41(3), 1–58. https://dl.acm.org/doi/10.1145/154880.1541882

de Vries, L. P., Baselmans, B. M. L., & Bartels, M. (2021). Smartphone-based ecological momentary assessment of well-being: A systematic review and recommendations for future studies. *Journal of Happiness Studies*, 22(5), 2361–2408. https://link.springer.com/10.1007/s10902-020-00324-7

Fricker, R. D., Hegler, B. L., & Dunfee, D. A. (2008). Comparing syndromic surveillance detection methods: EARS’ versus a CUSUM-based methodology. *Statistics in Medicine*, 27(17), 3407–3429. https://onlinelibrary.wiley.com/doi/10.1002/sim.3197

Hekler, E. B., Rivera, D. E., Martin, C. A., Phatak, S. S., Freigoun, M. T., Korinek, E., Klasnja, P., Adams, M. A., & Buman, M. P. (2018). Tutorial for using control systems engineering to optimize adaptive mobile health interventions. *Journal of Medical Internet Research*, 20(6), e214. http://www.jmir.org/2018/6/e214/

Hulme, W. J., Martin, G. P., Sperrin, M., Casson, A. J., Bucci, S., Lewis, S., & Peek, N. (2021). Adaptive symptom monitoring using hidden Markov models - An application in ecological momentary assessment. *IEEE Journal of Biomedical and Health Informatics*, 25(5), 1770–1780. https://ieeexplore.ieee.org/document/9226077/

Koizumi, D., Matsuda, T., & Sonoda, M. (2012). On the automatic detection algorithm of cross site scripting (xss) with the non-stationary bernoulli distribution. In *The 5th International Conference on Communications, Computers and Applications (MIC-CCA2012)* (pp. 131–135).

Liu, H., Xie, Q. W., & Lou, V. W. Q. (2019). Everyday social interactions and intra-individual variability in affect: A systematic review and meta-analysis of ecological momentary assessment studies. *Motivation and Emotion*, 43(2), 339–353.

MacDonald, S. W., Nyberg, L., & Bäckman, L. (2006). Intra-individual variability in behavior: Links to brain structure, neurotransmission and neuronal activity. *Trends in Neurosciences*, 29(8), 474–480. https://linkinghub.elsevier.com/retrieve/pii/S0166223606001251

Mohan, S. (2021). Exploring the role of common model of cognition in designing adaptive coaching interactions for health behavior change. *Journal of Interactive Web-based Data Visualization with R, plotly, and shiny, 1*(1), 1–30. https://dl.acm.org/doi/10.1145/3375790

Monacelli, G., Zhang, L., Schlee, W., Langguth, B., & Murphy, T. B. (2023). Adaptive data collection for intraindividual studies affected by adherence. *Biometrical Journal*, 65, 2200203. https://doi.org/10.1002/bimj.202200203

How to cite this article: Monacelli, G., Zhang, L., Schlee, W., Langguth, B., Ward, T. E., & Murphy, T. B. (2023). Adaptive data collection for intraindividual studies affected by adherence. *Biometrical Journal*, 65, 2200203.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.