RESEARCH

Characteristic Latent Features for Analyzing Digital Mental Health Interaction and Improved Explainability

Max-M Theilig[1,2*], Ashley A Knapp[2], Jennifer Nicholas[2,3,4], Rüdiger Zarnekow[1] and David C Mohr[2]

*Correspondence:
m.theilig@tu-berlin.de,
max.theilig@northwestern.edu
[1]Technische Universität Berlin, Chair of Information and Communication Management, Strasse des 17. Juni 135, 10623 Berlin, Germany
Full list of author information is available at the end of the article

Abstract

Background: Using smartphones and wearable sensor technology has sparked a broad engagement of data science and machine learning methods to leverage the complex, assorted amount of data. Despite verified processes, there is a reported underdevelopment of user engagement concepts, and the desire for high accuracy or significance has shown to lead to low explicability and irreproducibility. To overcome these issues, we aim to analyze principal characteristics of everyday behavior in digital mental health.

Methods: We generated five latent features based on previous research, expert opinions from digital mental health, and informed by data. The features were analyzed with descriptive statistics and data visualization. We carried out two rounds of evaluations with data from 12,400 users of IntelliCare, a mental health platform with 12 apps. First, we focused to proof concept and second, we assessed reproducibility by drawing conclusion from distribution differences. User data was drawn from both research trials and public deployment on Google Play.

Results: Our algorithms showed increased rationale for the basic usage of apps with different underlying behavioral strategies. Measures of the distribution of user's allocated attention, the user's circadian behavior, their consecutive commitment to a specific strategy, and users' interaction trajectory curve are perceived as transferable to the public data set. Because distributions between research trial and public deployment were similar, consistency was shown regarding the underlying behavioral strategies: psychoeducation and goal setting are used as a catalyst to overcome the users' primary obstacles, sleep hygiene is addressed most regularly, while regular emotional exposure is avoided. Relaxation as well as cognitive reframing have increased variance in commitment among public users, indicating the challenging nature of these apps. The relative course of the engagement (learning curve) is similar in research and public data.

Conclusions: The deliberate, a-priori engineered features were reproducible across app users from both data sets. These features led to improved results as well as increased interpretability, providing an increased understanding of how people engage with multiple mental health apps over time. Since we based the generation of features on generic interaction proxies, these methods are applicable to other cases in artificial intelligence and digital health.

Keywords: digital mental health; data science; feature engineering; latent features; mobile health; depression; anxiety

Background

The pervasiveness of mobile devices, wearable sensors, smart devices and social media is leading to the vast accumulation of individual user data. It is used to inform digital solutions that address individual health, including mental wellbeing and fitness self-optimization. While web analytics were an early attempt to demystify invisible use metrics and increase user engagement [1], there is a reported under-operationalization of, and lack of conceptual explanation for use within, eHealth technologies [2, 3]. The co-emerge of data science and machine learning as the state-of-the-art approach for quantifying and comparing available data sets [4] reinforce these efforts and call for appropriate methodologies to leverage digitally generated data for healthcare and digital epidemiology [3].

While at first sight, it seems processes such as automated machine learning provide an objective benchmark compared to other models leading to a reduced feature and model selection bias [5], this may only be the case for less complex data or a clear preference for prediction over explanation [6]. Unfortunately, with more complex data may come a greater number of features, increased chances of feigned significance or a lack of explicability. In sum, far from a silver bullet, machine learning is prone to errors when the input or the output evaluation is reverse engineered, and not thoroughly understood or explained [7, 8]. Therefore, it is possible to draw many conclusions from a given data set, leading to published findings that are not reproduceable or explainable, including the possibility of false positives [9].

The reliability of significant findings, the so-called replication crisis, is of particular importance, and while all fields related to quantification have this problem, the focus of discussion has been on its impact in psychology and medicine [10, 11]. Further, these areas have been a significant focus in digital health research; for instance, studies examining mental wellbeing and behavior in everyday life with data from mobile are at the forefront of digital health research [12, 13, 14]. In fact, follow-up studies often were unable to reproduce complex models and interdependencies in real-world context that consider app usage, events and communications log with their effect on user's individual well-being and their mood, e.g. ecological momentary assessment [15].

Given the commonly used screening tests, it is plausible how the field got to this point. As a solution, transparent reporting of key data and developing standardized characteristics is central to understand interventions and explain the way in which they engage and retain users [16].

Feature engineering as the hidden problem domain

To tackle the various inputs for artificial intelligence (AI) in scenarios of digital mental health (DMH) tools, usually simplifications and assumptions on a low level of data integration have to be made [17]. What to simplify and where to abbreviate the process is a multi-layered problem itself. Even for data that may appear quite trivial, feature engineering is considered a daunting and complex task [18]. At the system level, the question of how to accurately assess app use is raised, when objective measures, such as installation of an app, are not necessarily accurate [19]. For instance, studies suggest that real-world uptake of DMH apps, with a fluctuating range, is less than half compared to the reported values in research trials [16].

Being aware of aforesaid flaws, often researchers decide to go with any input mean that shows a minimum of variability to begin with. For instance, to use the frequency of app launches as a proxy variable for usage it might be required to have a stationary process or at least, it is beneficial to the chosen statistic model. Yet unfortunately, though not incorrect, these a-priori decisions often go hand in hand with more difficult interpretation and narrow system conditions, which has led to irreproducibility and a lack of credibility in research [9, 20]. Methods for the validation of various inputs, such as accelerometer data, have also shown to have systematic shortcomings within applications for medical machine learning [21]. It is complex to establish that causal link with the problem of the temporal modeling and controlling variables in research trials due to poor understanding of the used processes at hand [22]. Accordingly, only few researchers take the effort to do a distinct report with a comprehensive description of their implemented measures in addition to the primary outcomes of their work, e.g. as done by Taki et al. [23].

While the characteristics of digital interventions themselves are quite established and understood, there is still a lack of operationalization leading to a poor understanding of the link between users' uptake and intervention effect [24]. In sum, the relationship between research evidence and real-world uptake is key. There is a need to expand evaluation metrics and therefore, to improve efficiency in the feature engineering phase in a way that satisfies non-functional requirements such as interpretability [3, 18]. Otherwise, in light of the notion that efficacy trials largely emphasize internal validity over real-world issues, such as the technological environment, implementation, sustainability, and usage, the essential validation of novel treatment options may not be provided [25].

Objectives

To make a positive case for deliberate feature engineering instead of a brute force approach and to better understand use within research and real-world environments, we developed a set of latent temporal features describing user engagement for application in the digital mental health space. The aim of this study is to explore and establish features for AI use that indicate how users engage with a digital health app and investigate how they might increase our understanding of intended outcome replication in real-world settings.

The use of real-world behavioral data sets, which leverage a vast number of users, is considered a novel way to learn about mental health interventions [26]. A strength of our approach is our access to two similar data sets from research trials and public deployment and relying only on the identical features for comparison. We conclude by highlighting the major generalizable findings within this data, and outline future directions to reach an abstract set of reliable everyday life use features.

Methods

Our coherence driven method has two steps: First, we propose five generic features, justify their reasoning, give practical visualizations, and provide detailed algorithms for replication. Then, we show proof of concept by implementing these features, and discuss transferability and replication potential across two distinct types of app deployment, the control conditions of research trials and a public deployment

through the Google Play Store. We replicate the five features across these two data sets, analyze their distribution within those sets, and discuss the logical consistency of the proposed characteristics. We explain this in detail in the corresponding result section.

The IntelliCare platform

This work draws on the data gathered with, and the concepts implemented in, a large publicly accessible platform for mental health called IntelliCare. The IntelliCare app platform includes a hub app, that coordinates users' experience, and 12 clinical apps, each designed to target a specific behavioral or psychological treatment strategy (e.g. goal setting, behavioral activation, etc.). The clinical apps include: *WorryKnot*, *BoostMe*, *DayToDay*, *MyMantra*, *Aspire*, *DailyFeats*, *ThoughtChallenger*, *PurpleChill*, *MoveMe*, *SocialForce*, *iCope*, and *SlumberTime*. IntelliCare apps facilitate interactive skills training and are designed to be used over multiple short interactions. The hub app performs a gatekeeper function to the platform by consolidating notifications, providing recommendations of the clinical apps, presenting psychoeducation, and delivering surveys and questionnaires.

Across the apps, interaction time varies between a few seconds and some minutes, depending on the apps function. For example, apps designed for symptom monitoring or tracking goals are often used during multiple short spurts of time while apps designed for cognitive exercises or relaxation activities are used for longer. A full description of the apps can be found elsewhere [27]. The IntelliCare apps have been shown to be effective and efficacious at reducing symptoms of depression and anxiety among adults in two research trials [28, 27]. In the trials, all participants were expected to stay engaged with IntelliCare apps for 8 weeks and, were supported by a Coach who supported participants via text messaging to promote engagement.

The second component for the used data set comes from an open, public deployment of IntelliCare on Google Play [29]. The public deployment of IntelliCare has evolved largely congruent to the research trials and has been accessible since September 2014. Some changes occurred in the app suite over time. For example, modifications to apps based on user feedback were made, apps were added (e.g. *BoostMe* [30] and one app was removed (*MeLocate*) as it did not perform as expected.

Proposed interaction features

Below we outline five app use features that we believe represent interaction metrics of app engagement based on the described data sets, discussions with experts from the field, and our research to date [31, 32].

Feature 1: Adjusted Usage

Assessing app usage is a common method of evaluating behavioral engagement with an app and often the basis for further detailed assessments [17]. The measure of app use proposed here provides additional information beyond that of isolated count of events or duration. It captures the amount of app events in relation to time spent, for each person and in relation to the whole population. Such app interaction data are commonly measured as a sequence of events (i.e., each time-stamped interaction

that a person has with an app, like pressing a button). Therefore, app sessions are defined as the time from first to last event, separated by a fixed number of minutes before the next session starts. However, this session length typically does not account for the different levels of interaction that can be performed among multiple sessions of one app (e.g., putting the app into background to check a push notification in one occasional instance) and between different apps (e.g., reading each item description in one app; hectically clicking through a checklist in another app).

To compare usage, this variability has to be adjusted for each app type, user, and session, as depicted in Fig. 1. On a global level, the average event count and duration represents the typical requirements for a session to be completed within each app type, which is implicitly determined by the underlying task and app programming. On an individual level, users vary in the speed they work through an app session. For example, one individual may tend to take their time and consider every step, while another may work rapidly. Further, individual users may vary over time in the length of time they spend on similar tasks within any app, e.g. when they are in a hurry. This process is depicted in Algorithm 1. First, the session is adjusted with respect to the global duration and event count characteristic for that app type. And second, the length is adjusted for each user with respect to the user's individual pace of interacting with that app.

Algorithm 1 Pseudocode to calculate Adjusted Usage. First, the average global weights are calculated for each distinct app. Second, for each user, the individual weights for their distinct app sessions are set in relation to the global weight. That factor is used to adjust the duration of each session.

for all $k \in \mathcal{A}pps$ **do** ▷ Global level
  $n = |\{s \in \mathcal{S}essions_k\}|$
  $GblWeight(k) = \frac{1}{n} \sum_{s \in \mathcal{S}essions_k} s.\textsc{Duration}$
end for
for all $i \in \mathcal{U}sers$ **do** ▷ Individual level
  $n_k = |\{s \in \mathcal{S}essions_k(i)\}|$
  $UserWeight_k(i) = \frac{1}{n_k} \sum_{s \in \mathcal{S}essions_k(i)} s.\textsc{Duration}$
  $IndvFactor_k(i) = {}^{UserWeight_k(i)}/_{GblWeight_k(i)}$
  for all $s \in \mathcal{S}essions(i)$ **do**
    $AdjUsage_{i,k}(s) = IndvFactor_k(i) \cdot s.\textsc{Duration}$
  end for
end for

The *Adjusted Usage* feature measures the length of time an app is used adjusted in a two-stage process. Therefore, this feature now carries information about the personal habits of each user, for all app types and compared to the whole population. In comparison to other methods, more intense use over a consistent period of time or a longer duration at an equal level pace of use will result in a higher value of *Adjusted Usage*.

Feature 2: Allocated Attention

The *Allocated Attention* feature proposed here captures the overall amount of usage for each app, normalized by the most used app, in one single rational number. It provides a higher granularity measure by comparing attention across apps, and therefore a promising alternative for apps downloaded. Different apps within a suite

of apps usually offer distinct purposes, such as different behavioral strategies in a digital mental health platform, to the user. Accordingly, assessing which apps are used most compared to other apps is an important metric to capture. This has typically been measured by the discrete number of downloaded apps. This measurement, however, does not account for how much attention is dedicated to the different apps after being downloaded, skewing the comparison of app attention among users. Thus, the intended ability to distinguish between users which used the most different apps and those that used least or none at all after download is not captured by an apps-downloaded metric.

When there is an explicit definition of usage, e.g. the aforementioned *Adjusted Usage* or a count of events, it is possible to rank the apps according to the sum of this parameter. Then, in reference to the most used app, all other apps will only receive a fraction of the user's attention, when looked at over the user's whole lifetime of app use. For example, in Fig. 2, aside the favorite app (App 1), the user attended to two more apps (App 2 and App 3) with about half of their resources dedicated to each of those two apps. Therefore, the achieved score of 2.2 is more representative than accounting for 5 downloaded apps, of which two have been abandoned (App 4 and App 5).

Algorithm 2 Pseudocode to calculate Allocated Attention. After initialization, the overall app usages for each user have to be sorted with respect to different app types. Their proportions relative to the favorite app are summed up.

Initialize:
  $Att(i) \leftarrow 0,\ \forall i \in \mathcal{U}sers$
  $OverallUsage_i[\ ] \leftarrow \sum_{\mathcal{S}essions_i} \mathcal{U}sage_i(k)\ \ \forall k \in \mathcal{A}pps$ ▷ Index, app accessible frame
for all $i \in \mathcal{U}sers$ **do**
  $OverallUsage_i[\ ] = \text{SORT\_ASCENDING}(OverallUsage_i[\ ])$
  for $k \leftarrow 0$ to $AppUse_i[\ ].\text{LENGTH} - 1$ **do**
    $RelAtt_i(k) = {}^{AppUse_i[k]}/_{AppUse_i[0]}$ ▷ Usage compared to favorite app
    $Att(i) = Att(i) + RelAtt_i(k)$
  end for
end for

The *Allocated Attention* feature represents the number of apps users allocated their attention to, over their whole lifetime in the respective ecosystem. For each user, the features ranges from 1 to the absolute number of downloaded apps as its supremum (in Fig. 2 that would be 5). A higher value indicates that the user's attention is spread across several different apps. This metric is an alternative for the number of downloaded apps, in a mathematically continuous sense, and gives a much clearer view of how many apps an individual dedicated their time to.

Feature 3: Circadian Use

The *Circadian Use* feature measures the degree to which a person tends to use certain apps at the same time each day. The consistent use of an app is likely the result of how well the user incorporates the use of the app into their daily lives, forming a habit that may be represented through a temporal use pattern. Since these patterns may vary across users, depending on their preferences, lifestyle, or work schedules, it is not sufficient to look at the same time of the day for all users. For example, users may variously prefer to use an app in the morning, throughout

the today, or mainly in the evening. In sum, we aim to capture this variance between users.

Therefore, the *Circadian Use* feature focuses on deviations from past use, with a lower diversion depicting daily use of an app at similar times every day. Fig. 3 depicts the implied circadian deviations for one 24-hour period. When a specific app session is known it is simple to go back 24 hours and look for similar app activities in a fixed time window. With similar use sessions identified, the *Circadian Use* can be calculated by summing up the differences, to the reference time, within the past 24-hours. While the fixed window in which to identify similar app activities should be confined (smaller than 12 hours), a similar logic can then be applied to more days in the past, i.e. past 48 hours, past 72 hours, and so on. Looking back further than 24 hours is required as it is not assumed that a user has a daily rhythm. To account for such patterns, we suggest the unassuming way of averaging the previous days with a decaying effect by dividing by the integer number of days passed.

Algorithm 3 Pseudo code to calculate Circadian Use. For each user and each session, all sessions within a 6-hour time window are gather for a defined number of days in the past. The root mean square of their deviations is summed up in a discounting way and finally, averaged for a user-representative measure.

Initialize:
  $\mathcal{D}ays \leftarrow \{1, 2, 3, 4, 5, 6, 7\}$ ▷ E.g. one week
  $Sessions_{d,k}[\ ] \leftarrow \mathcal{S}essions(i)\ \ \forall d \in \mathcal{D}ays, \forall k \in \mathcal{A}pps$
  ▷ Gather timestamp and app properties
for all $i \in \mathcal{U}sers$ **do**
  for all $s \in \mathcal{S}essions(i)$ **do**
    $PrevSessions_d(s)[\ ] \leftarrow Sessions_{d,k}(s)$
      $\forall prev \in \mathcal{S}essions_i \mid \text{ABS}(prev.\text{TIME} - (s.\text{TIME} - d \cdot 24\,\text{h})) < 6\,\text{h}$
      ▷ Assign the set of (all) previous sessions
    $PrevDeviations_{d,k}(s)[\ ] \leftarrow (prev.\text{TIME} - (s.\text{TIME} - d \cdot 24\,\text{h}))$
      $\forall prev \in \mathcal{P}revSessions_d, (s) \mid prev.\text{APP} = s.\text{APP}$
      ▷ Assign the set of deviations for each reference session
    $DayFid_d(s) = \text{ROOT\_MEAN\_SQUARE}(PrevDeviations_{d,k}(s)[\ ])$
    $CircUse(s) = \sum_{d \in \mathcal{D}ays} \text{IF\_ZERO}(12, DayFid_d(s)) \cdot \frac{1}{d}$ ▷ Mitigate by days passed
  end for
  $n_k = |\{s \in \mathcal{S}essions_k(i)\}|$
  $AvgCircUse_k(i) = \frac{1}{n_k} \sum_{s \in \mathcal{S}essions_k(i)} CircUse(s)$
end for

The *Circadian Use* feature depicts the average time window in which the user engages with the app or platform daily. Lower numbers represent integration of app use into a person's daily life. For example, the measure of 0 would be a person using an app or platform at the exact same time every day. A measure of 1 means the user integrates the daily use of the app within a one-hour window of time. A high value of $CircUse = 4.12$, as depicted in Fig. 3, implies a daily use but within a larger time window of about 4 hours. The measure is limited to window of 12 hours, since larger values refer to the day before and cannot therefore be considered circadian.

Feature 4: Follow-Up Commitment

The *Follow-up Commitment* feature captures the degree to which any person has shown a tendency of consecutive, repetitive use of an app in the past. In an ecosys-

tem of apps, people may use several apps concurrently during a given time period. Some apps may be intended to be used on an as-needed basis and do not require frequent and consecutive usage, while others may be designed such that each session builds on the last, requiring consistent usage. Indeed, in digital interventions it might be of interest to learn skills and continue to practice those learned skills via using apps. In contrast, it is also possible that once a skill is learned there is not a need to revisit that app. Because of this ambiguity, it is critical to look at a user's commitment to an app.

The *Follow-Up Commitment* feature measures this behavior. In Fig. 4 the label "+1" indicates that the same app was used consecutively, whereas the label "×" indicates a change, in that a new app was used next. If the constraint of looking at one specific app is dropped, a more general commitment level for each user can be calculated. Fig. 4 visualizes this procedure for App 1 commitment in the upper row, and for the general commitment (across all apps) in the lower row, here the re-occurrence of App 1 and App 2. Therefore, Algorithm 5 describes how the user's sequential pattern of use can provide useful information about the commitment. It sorts the sequence of apps used by an individual and counts consecutive occurrences of the same app. Finally, a relative factor of commitment in relation to total count of app sessions is created, a value between 0 and 1.

Algorithm 4 Pseudo code to calculate Follow-op Commitment. First, the current lifetime sequence of app sessions for each user is sorted, including all distinct app types. Only consecutive use of the same app is counted and put in relation to its respective overall session count.

Initialize:
$ConsCount_k(i) \leftarrow 0 \quad \forall k \in \mathcal{A}pps$

for all $i \in \mathcal{U}sers$ **do**
  $SessSeq_i[\ ] = \text{SORT\_ASCENDING}(Sessions(i))$ ▷ Put sessions in a sorted frame
  for $w \leftarrow 1$ to $SessSeq_i[\ ].\text{LENGTH} - 1$ **do**
    if $\text{GET\_APP}(SessSeq_i[w-1]) = \text{GET\_APP}(SessSeq_i[w])$ **then**
      $ConsCount_k(i) = ConsCount_k(i) + 1$
    end if
  end for
  $n_k = |\{s \in \mathcal{S}essions_k(i)\}|$
  $FollowCom_k(i) = \frac{1}{n_k} \cdot ConsCount_k(i)$
end for

The *Follow-Up Commitment* feature captures how consistently a person comes back to using the same app. This feature is an indicator of the underlying engagement with, or commitment to that one app. A high value close to 1 represents consecutive use of or commitment to a single app, while a value close to 0 represents an increased alternation between multiple apps. Combining this feature with the previously introduced feature of *Allocated Attention* leads to an even more real-world explanation: When a user evenly attends to multiple apps, therefore showing low commitment to these apps, this is similar to multitasking.

Feature 5: App Use Trajectory

The *App Use Trajectory* feature provides a measure of the time spent on an app in relation to the time since it was first opened. This reflects the time period during which a person is most likely to be engaged with an app, which is often not consistent

over time. In the beginning, users might increase engagement as they get more comfortable or familiar with the app. Then, after people steadily use an app with some frequency, that activity is expected to fall off over time [33]. However, within that process there may be many trajectories, as depicted in Fig. 5. Tracking the distinct trajectory of app usage may be useful to assess the quality of app design or to predict how that individual will respond to other apps [34].

This feature indicates the time spent on an app in relation to the time passed since the app was first used. At one extreme would be a user that opens the app just once, days after download. At the other extreme would be a user that constantly increases their time spent on the app in the immediate days after the initial download. To calculate an average for a user, a fixed time window is necessary to uphold comparability, for example, the first two weeks.

Algorithm 5 Pseudo code to calculate App Use Trajectory. Initially, for each user the sessions are sorted, and the timestamps of download for used apps are identified. The accumulated session durations, discounted by the time passed, are put into a temporary array for each distinct app. Finally, this array is averaged into a score for each app.

for all $i \in \mathcal{U}sers$ **do**
  $SessSeq_i[\ ] = \text{SORT\_ASCENDING}(Sessions(i))$ ▷ Put sessions in a sorted frame
  Initialize:
  $time_0(k) \leftarrow \text{GET\_TIME}(SessSeq_i[0]) \quad \forall k \in \mathcal{A}pps$ ▷ Initial download
  $DurSum_k(i) \leftarrow 0 \quad \forall k \in \mathcal{A}pps$ ▷ Variable for app time used
  for $w \leftarrow 1$ to $SessSeq_i[\ ].\text{LENGTH} - 1$ **do**
    $DurSum_k(i) = DurSum_k(i) + \text{GET\_DURATION}(SessSeq_i[w])$
    $time\_passed_k = \text{GET\_TIME}(SessSeq_i[w]) - time_0(k)$
    $UseTraj_k(i)[w] = {}^{DurSum_k(i)}/_{time\_passed_k}$
  end for
  $AvgUseTraj_k(i) = \text{AVERAGE}(UseTraj_k(i)[\ ])$
  ▷ Preferred measure of central tendency, e.g. median
end for

The *App Use Trajectory* feature provides a measure of when a user most strongly engages with an app after being introduced to that app. When the coefficient increases over time, it indicates that the user displays higher interest and use, while a constant level over time indicates stable interest and use, and a decreasing coefficient reflects lower interest and use.

Results

We carried out two rounds of evaluations with the afore described collection of features. First, we ran the calculations for the users from the research trials (RT) to test for proof of concept and plausibility. And second, we evaluated the calculations from the public deployment (PD) compared to the RT to assess reproducibility. Primarily, we focused on plausibility for both runs regarding the behavioral strategy behind each app, and in a second level we drew conclusions from possible differences between the research and public users. To visualize differences and similarities in the distribution of the replicated features, we will largely rely on the Letter-Value Plots [35], which are especially suited to compare data similar to our case. We start by describing the data subsets as well as their statistic properties and constraints in more detail.

General user statistics

Data from 400 RT participants were examined. We excluded early dropouts that were active less than 10 days and users showing specific artifacts such as performing all of their interaction on the very last day of their trial, resulting in 386 users. Participant selection and support in these trials resulted in more consistent app use, with user-initiated events of more than 1.2 million events from ca. 140k app sessions. The typical users from the RT were classified with an average (median) lifetime of $\bar{x} = 244$ ($\hat{x} = 179$) days (note, app use data following the 8-week trial was included as most participants continued using the apps), $\bar{x} = 6406$ ($\hat{x} = 2367$) interaction events, and an activity of $\bar{x} = 28.4$ ($\hat{x} = 15$) events per days.

At the time of this study, 27,241 PD users had issued more than 85,000 app downloads from the IntelliCare suite from the Google Play Store. Ten of 12 apps had received a rating between 3 and 4 stars (out of 5 stars). For this work we excluded particular unpopular apps (Social Force and iCope). Simultaneously, the underlying database carried more than 45 million events, which was about 73 GB of data. As is common in public deployment for freely available apps, some users were not consistently engaged. Hence, we excluded 15,218 (55.8% of total) users, whose data were too sparse to allow for feature calculations. Explicitly, users with at least 30 events overall, more than 7 days of app use within the app suite, and an activity of at least 1 interaction every 10 days, were included. This resulted in 12,023 users, for which the data was sufficient to include, as feature calculations made up for a volume of around 6.3 million user-initiated events resulting in about 500k sessions. The PD users showed an average (median) lifetime of $\bar{x} = 137$ ($\hat{x} = 58$) days, $\bar{x} = 2152$ ($\hat{x} = 142$) interaction events, and an activity of $\bar{x} = 16.8$ ($\hat{x} = 3$) events per days.

Adjusted Usage reproducibility

The Adjusted Usage as described and its correlation to multiple broad use indicators are shown in Table 1, for the RT as well as for the PD. Most interestingly, it might be compared to the 2nd column *Sessions* count, which is the common, most naive, indicator for number of app sessions (app launches).

Comparing the *Usage* to the *Sessions*, values seem to be close for the most part in the RT and are almost identical in the PD. For the RT, usage shows improved correlation values to the count of Events ($r = 0.91$ over $r = 0.84$), the *Time* engaged ($r = 0.98$ over $r = 0.73$), and the *Frequency* of engagement ($r = 0.63$ over $r = 0.42$). This shows major advantages of the Usage parameter since it more broadly captures the timely, dedication aspects. That holds true even for comparable smaller sample sizes, like RT, and therefore improves explainability.

When comparing the RT and PD triangular matrices (Table 1), we find that reproducibility is given for most of these basic parameters, strong effects ($r \geq 0.65$) get stronger, weak effects get weaker ($r \leq 0.35$), overall significance, in terms of p-values, is increasing slightly (not depicted). Interestingly, the coefficients in the column of *Frequency* of engagement seems to be the exception to that rule. On a second thought, this could be a consequence of the comparably short time of the RT. Since the trial length was fixed (8 weeks) and the majority kept close to that, the denominator involved a narrower corridor. Therefore, the frequency was

largely tied to the numerator (events) and created a shadow correlation. For the low commitment nature of the PD, with unlimited time, a lot of events (clicks) in a certain window of time are not per se related to many events, sessions, or increased usage overall. Even more so, the commonly required Frequency is given by behavioral strategy behind the app, the implementation of the developer, and personal mental processing preferences. That again is in line with the argumentation from the subsection for *Adjusted Usage*. Negative effect strengths ($r \ll 0$) was found for Lifetime and Frequency for RT as well as PD. Noteworthy, Frequency is not negatively correlated to the Time of Engagement, which is in line with the above argumentation and was already considered constructing the Usage parameter. Strengthening that point of intrinsic motivation, the *Apps* downloaded correlation is considerably higher for the RT. That means, if an encouraged user decided to download an app, they actually intended to use it. On the other hand, users from the PD were motivated to look to try new apps, therefore presumably lacking intrinsic drive to stick with something that did not immediately catch their interest.

Insignificant coefficients were not to be noted for both populations. Lowest significance, still at an alpha level of 5%, for the PD is given for the relations of days of *Lifetime* and the *Apps* downloaded. This is plausible since all apps were available from the beginning and therefore, when the p-values are increased correlation is low, or very low as for the PD (see * and ** in Table 1).

Allocated Attention by favorite apps

While many users of the RT downloaded all apps (11), the allocated attention distributed by users is much lower peaking at around $Att(i) = 6.5$ for the RT as well of for the PD. We depict more detailed differences between the RT and PD data by grouping the app attention according to the user's favorite app, in Fig. 6.

On average (median), the *Allocated Attention* parameter is about $\bar{Att} = 2.78$ ($\hat{Att} = 2.64$) for the RT users and about $\bar{Att} = 1.85$ ($\hat{Att} = 1.64$) for the PD users. That means RT users devoted about 64% of the attention to apps other than their favorite, while average PD users spent a slight minority of about 46% with apps other than their main one. Simply said users from the PD focused on fewer apps that catered to their needs, in average 1 app less. This also holds when looking at the lower quartile of users where 53% ($Att(i) = 2.11$) of usage was dedicated for the favorite app, leaving some attention to also follow different behavioral strategies, for the RT. While at least 97% ($Att(i) = 1.03$) of attention went into focusing on their primary app, for the lower quartile PD users. This seems plausible since probably they were not even aware that a whole suite of apps exists. Nonetheless, it is safe to say a majority of the users from the PD used more than one app a significant amount. Therefore, that primary goal of the clinical study replicated well among PD users.

Looking at specific apps, we find that the broad picture of the RT is reproduced for the PD. DayToDay and DailyFeats have the lowest scoring distributions in the RT and show the minimum for the PD users (compare medians in Fig. 6). This can be explained by the behavioral strategies behind these apps. Psychoeducation (DayToDay) and goal setting (DailyFeats) train the more general ability to cope with psychological obstacles. Hence, the need for another app that addresses a different

behavioral problem is reduced, when these strategies are internalized. *MoveMe* on the other hand has a high scoring distribution for RT and plausibly, also is the highest scoring for the PD users, because it is demanding in a timely sense by design. Physical exercise activation helps to overcome amotivation but does not allow to address a problem itself. Therefore, a user has to incorporate at least one additional goal, making it the only primary app that scores a median App Attention below 50% ($Att(i) \approx 2.05$) in the PD. Due to the complex nature behind MyMantra (create an encouraging virtual album), this app requires extensive engagement, which was facilitated in the RT environment but not in the PD. Hence, the MyMantra median score shows the biggest gap, meaning that it is least reproducible among the PD users.

Circadian Use characteristics

For the calculation of the circadian behavior within each app (e.g. App 1, upper case in Fig. 3), about 106k sessions from the RT and about 395k sessions for the PD qualified, due to the constraints of coming back for using the same app at all. Detailed differences between the RT and PD user behavior regarding circadian interaction within each app are displayed in Fig. 7.

On average (median), users came back in a time window of $Avg\bar{C}ircUse = 16.3$ ($Avg\hat{C}ircUse = 15$) hours for the RT users and within $Avg\bar{C}ircUse = 17.7$ ($Avg\hat{C}ircUse = 16.9$) hours for the PD users, cumulated for the last week. Assuming linearity, that would imply an average time window of about 2.3 hours on a daily basis, 2.5 hours for the PD respectively. Following our proposed scale on the other hand, this is well above the 12 hours time window, which was defined as circadian behavior on a one-day basis (compare Fig. 3). Therefore, average users do not show circadian behavior but skipped at least one day for the RT, and almost two days for the PD (18 hours mark). The best scoring quartile was with $AvgCircUse = 7.9\,\text{h}$ below the 12-hour barrier for the RT, $AvgCircUse = 8.8\,\text{h}$ for the PD accordingly. Noteworthy, the worst scoring quartile of the PD was even at $AvgCircUse = 28.3\,\text{h}$, somewhere between a 5-Day and a 6-Day rhythm. That means they barely managed to come back using the same app once a week. In general, this behavior seems plausible since the users from the research trial were motivated to interact regularly and received motivational text messages from coaches. Users from the PD were not stimulated by coaching as well as more likely to abandon the regular usage for a longer period of time.

For the individual apps, we find that *DayToDay*, *DailyFeats*, and *SlumberTime* show the best median circadian user behavior (below the comparable Overall median). For SlumberTime, this is quite conceivable since this app is supposed to support sleep hygiene and sleep happens comparably regularly within both populations. As for *DayToDay* and *DailyFeats*, there is evidence that psychoeducation and goal setting are also well integrated into the everyday life. *WorryKnot*, and *Aspire* are among the worst scoring apps, since their lower quartile is at the 12-hour barrier (no daily use) and their upper quartile is at the maximum value, for RT and PD alike. That points to the interpretation that the need for self-management strategies, regarding worries and values, appears at various points in time and very unregularly throughout the week.

Users' Follow-Up Commitment

For the calculation of the users' *Follow-Up Commitment* for each distinct app (e.g. App 1, upper case in Fig. 4), we were able to utilize 2186 user-app combinations from the RT and about 8335 user-app combinations for the PD. Due to fact that a user at least had to use two apps alternatingly for this parameter to be computable, a significantly smaller yield from the PD users was possible. As visualized in Fig. 8, the distributions are diverse.

On average (median), users from the RT shows a consecutive commitment of only $\bar{FollowCom} = 4.4\%$ ($\hat{FollowCom} = 1.6\%$), which is only about a quarter of what the PD characterized with $\bar{FollowCom} = 22.6\%$ ($\hat{FollowCom} = 5.8\%$). Thus, while an average user from the PD used the exact same app as before in one of five instances, for the RT this was the case in one of 20 instances. On the one hand, this is quite plausible hence the RT on average utilized more apps than the PD (see above). On the other hand, the factor for *Allocated Attention* was well below 2 for the PD as we found above, and here it is close to 5. Therefore, when the RT uses not even twice as much apps, but is only committed to a fifth of what the PD is, that points in the direction that users employ some kind of multi-tasking strategy. They distribute their attention across multiple apps, and develop some rotation, therefore they are not consecutively committed. Another noteworthy addition, the interquartile range (IQR) for the PD users, is significantly higher for the PD with 25%, compared to 3.7% for the RT. Intuitively, this follows the more disparate nature of the PD users and their usage strategies.

Besides the big increases of the IQR when comparing RT and PD, the median values for *DailyFeats* and *DayToDay*, about $FollowCom_k(i) \approx 15\%$ for PD, are among the highest. In contrast, it is marginally small for *MoveMe*, *Aspire*, *WorryKnot* and *BoostMe*, with about $FollowCom_k(i) \approx 2\%$ for the PD. That is equally as low as for the RT and is explained by the logic that people might repetitively want to improve their knowledge through psychoeducation with *DayToDay*. For apps such as *MoveMe* or *Aspire*, however, it would not be plausible to consecutively increase your motivation or resolve negative thoughts, without actually doing an activity in between for what you have been motivating yourself.

App Use Trajectory influence

According to the algorithm we described, every user may reach their individual high of engagement at a different moment in time (e.g. 2nd Day in Fig. 5). We find, that most users are comparably engaged between the days 8 to 12 after their initial download, with 6210 sessions for the RT and 20412 for the PD. The App Use Trajectory distribution for sessions occurred within that window are depicted in Fig. 9, grouped by the corresponding app.

On average (median), RT as well as PD users spend about $\bar{AvgUseTraj} = 2.5$ ($\hat{AvgUseTraj} = 1.7$) minutes per day passed during that comparable time period. Compared to the time window before (day 0 to 8, not depicted), with $\bar{AvgUseTraj} = 8.9$ ($\hat{AvgUseTraj} = 3.7$) minutes per day passed for RT and even $\bar{AvgUseTraj} = 12.6$ ($\hat{AvgUseTraj} = 4.2$) minutes for the PD users, this is a significant reduction. But put in perspective to the overall average session duration of only $\bar{AvgUseTraj} = 73$ ($\hat{AvgUseTraj} = 15$) seconds, this still qualifies as growing

or constant phase of interest. When comparing the RT to the PD regarding their learning curve in general, differences are marginal.

Looking at specific apps, *PurpleChill* and *Slumbertime* show the highest App Use Trajectory score. Therefore, we assume that our initial hypotheses, specifically that users strongly engage in the beginning, might not be as accurate for those apps. Since those apps focus on relaxation and sleep hygiene, they simply offer activities that require more time. For instance, these two apps offer audio recordings to facilitate relaxation, which might no longer be necessary once a regular habit of relaxation is reached. Therefore, the learning curve is very discrete. That artifact aside, *Thoughtchallenger* and *WorryKnot* can be seen as the runner ups, which might be related to the training it takes to relax and unwind. *MoveMe* and *Aspire* seem to require the least engagement in the beginning according to the App Use Trajectory. It is plausible that the concept of goal setting and motivational exercises are quite well known and therefore are characterized by a shallow learning curve.

Discussion

Principal results

This study evaluated five app use features to explain the user engagement over time with the 12 apps from the mental health platform of IntelliCare. These features are formalizable, address plausible behavioral app use patterns, and are reproducible across users in research trials, most of whom had a human coach to support engagement, and general public users who found the apps on an app store and used them on their own. The notion of these features was designed to have an improved tangibility and increased rationale relative to common engagement concepts as of today. Shown reproducibly across different or user groups suggests that such features within AI can be used to better understand user behavior for app choices on a platform. Consequently, our approach can be employed in AI methods to overcome replication issues and to produce personalized algorithms for app choices on an app platform that are likely to produce greater levels of engagement, which should in turn result in improved well-being.

Feature 1: Adjusted Usage

The *Adjusted Usage* parameter is the fundamental use metric for app engagement in this work, measuring the amount of app events in relation to time spent, for each individual but in perspective to the whole population.

The brain is always searching for ways to retain energy, thereby choosing the path that requires the least mental energy [36]. This self-regulation is central to effective human functioning and is expressed in meeting personal standards or achieving certain goals [37], here e.g. “To achieve my goals I will try to work with these apps”, for the overarching objective, and “I will concentrate on this app for now”, for the recent interactions. Hence, users focus on what is most important and regulate the interaction to comply with their short- and long-term interests [38]. The goals on the other hand are given by each app and users choose what caters best to their needs. Therefore, individual values, available mental energy and goals of interest call for two levels of adjustment as we implemented them above on the *Global Level*,

for each distinct app type, and on the *Individual Level* to account for users work-through speed. We show that this is feasible, and the increased plausibility is also reproducible in public environments.

This provides richer information than the isolated the count of clicks or duration, increasing the explicability of DMH interaction. Not surprisingly, this metric is highly correlated with other commonly used metrics such as number of app sessions or time on an app, as it incorporates information from both. While this metric is comprehensive, it is not specific, as simple usage metrics can reflect a variety of psychological states, such as motivation, fit of the app to personal goals, intention to use. To obtain more specificity in user behavior, we created 4 additional features.

Feature 2: Allocated Attention

The *Allocated Attention* quantifies the overall amount of use, normalized by the most used app, in a single rational number. It provides a higher granularity measure for apps downloaded. People tend to use many apps, not just one. Accordingly, DMH should no longer view the use of a single app, rather the use of multiple apps over time, that can be embedded in a broader online platform or ecosystem. Viewed from that perspective, the relationship of the use of one app relative to other apps can provide potentially valuable information. For example, some people tend to focus on one app at a time, a sort of time-limited brand loyalty [1]. These preferences can change over time. Very few people continue using a mental health app for more than two weeks [33]. It is possible that users may become bored with the app, or that the has served its purpose and the user is satisfied with the benefit gained. Among people who tend to prefer one app over others, this metric, over time, can reflect those shifts.

Some people tended to use several apps at the same time. This suggests that the individual has greater capacity to focus on more than one psychological strategy on a daily basis. This focus on multiple apps or strategies may simply reflect a sort of parallel goal pursuance, but not psychological multitasking. However, it is also possible that people find a generative or synergistic effect between the apps. For example, someone with insomnia might find that the insomnia app is enhanced by use of apps that offer relaxation exercises or emotion regulation techniques. To our surprise, the effect of distributing attention across apps on long-term engagement frequency was found to be somewhat higher for public deployment users, who just found the apps on the app store, compared to those who were enrolled in research trials of IntelliCare. This may reflect this synergistic support that apps may have for each other, which would not be as apparent for users who were also receiving some support and direction from a coach. Therefore, expanding existing apps with content to cross-utilize might be preferred over chasing short team user awareness by adding a new app to the platform.

Feature 3: Circadian Use

The *Circadian Use* measures the degree to which a person tends to use the apps at the same time each day. This is a measure comparable to regular recurrence on web pages or in social media, e.g. recency and loyalty indices [1], though they are not explained on a day-to-day basis. Especially for mental well-being, disruptions of

circadian rhythm have shown to modulate mood, attention and to promote severe disorders like depression [39]. We believe *Circadian Use* is a suitable marker of how well integrated apps are into the rhythms of a person's daily life.

It is well known that medication adherence is higher among people who are able to integrate pill taking into daily habits, such as placing medications next to the toothbrush as a reminder to take them in the morning or evening [40]. Indeed, a similar circadian metric of periodicity has been shown to be predictive sustained engagement with a web-based DMH intervention for depression [41]. High Circadian Use may not always be important. For example, while high Circadian Use is associated with engagement for many apps (e.g. consistent evening use of an app for insomnia predicts longer term engagement), this is not true of all apps. In particular, for those that are designed to be used on an as-needed basis, such as managing bouts of worry, Circadian Use is not strongly related to engagement. Thus, the utility of these metrics must be considered in light of the aims and strategies employed by any individual app.

Feature 4: Follow-Up Commitment

Follow-up Commitment refers to the degree to which an individual has shown a tendency in the past for repeated, consecutive use of the same app. This recurrent behavior which is acquired through frequent repetition [36], can also be interpreted as repetitive commitment that enforces people to keep doing as they always do [42]. In other words, this metric evaluates and implies past use as an indicator of future use. The idea that past behavior is the best predictor of future behavior is a well-establish principle in behavioral science [43]. However, while this is a strong principle, it cannot be universally applied, as some apps may provide benefit rather quickly, while other apps can require sustained engagement.

It is also likely that the context in which people are using any individual app can affect *Follow-up Commitment*. Users who came to the public deployment generally had higher *Follow-up Commitment* scores, compared to those in research trials. This is likely because most public deployment users only downloaded 1 to 3 apps, while users in the research trials received periodic recommendations to try new apps. Thus, follow-up metrics can be lowered by "interference" from other apps, that are competing for a finite amount of user attention. This type of interference, or encouraging a user to change their attention may be in the user's interest and is in line with finding that even the extensive use of multiple behavioral strategies might be related to greater effect sizes [44]. Thus, the interpretation of this metric should consider the context in which the app is being used, and whether interference is productive or unproductive.

Feature 5: App Use Trajectory

The *App Use Trajectory* measures the time spent on an app in relation to the time since it was first opened. This reflects the time period during which a person is most likely to be engaged with an app. While previous research has shown there are influential cues in the process of app search, selection and adoption [45], the literature on adaption and actual engagement after the installation is scarce.

We began this work expecting initially high levels of engagement that decline over time, consistent with existing literature [33]. However, we found the peak activity

occurred somewhere in the second week after the download, rather than in the very beginning, which was followed by the expected decline in interest over time. While it is unclear why our findings differ from the broader literature, we speculate that this may be reflective of a pattern when people are using an app platform, rather than a single app, where users might download several apps, select one for use, and then turn to others at a later time.

Limitations

Due to the considerable amount of data from constrained and public environments, a number of limitations may be mentioned.

First, while having a large real-world data set is key to investigate scientific transferability, it comes with a lot of artifacts of unknown origin. Despite looking at thousands of users as described, about half of the set had to be cleaned out due to the fast paced and low-commitment nature of the public app store as well as the fluctuation of app ranking and suggestions [46]. Additionally, when the data was sufficient to include for the initial feature engineering, potentially unconsidered external events may have skewed the data. Users from the public deployment may have been on vacation, changed their phones, or experienced hardware issues, just to mention a few. In general, users from the public deployment are likely more heterogeneous regarding their mental health condition, severity, and possible comorbidities relative to users in research trials. This could also be viewed as a strength in terms of a comprehensive data set.

Second, as we have described above, the interpretation of the features is dependent on the context. In the same way, the design of the features is dependent to some degree on the type of data, data granularity, and data quality that is utilized. While we believe we have identified features reflecting basic principles, these features would likely need to be adjusted, or at least debugged, to fit other contexts or data sets with weaker data quality compared to the data at hand.

Future work

The features developed here are designed in a generic way so that they might be transferred to other use cases, although as noted above, some modifications could be required. The range of predictive capabilities imply that only few changes are needed to transfer those features on further digital mental health apps and investigate primary outcomes. Future work could consider developing a feature store or repository, as a specific instance of knowledge management. Within that, features for a certain field (e.g. DMH) are being stored according to defined rules. Along with that, one should store supplementary information like if there is an indication of reproducibility, or how well they performed in a machine learning algorithm.

The features we developed only used app event data. Increasingly health apps are using sensing data from phone sensors and wearables, which could provide far greater information about the implementation of behavioral strategies in the context of individuals lives. For example, GPS data can provide features reflecting the frequency and patterns of movement that have been shown to predict depression and are related to behavioral activation strategies [47, 48]. Semantic location (the types of places visited) can estimate the time spent at home, a café, the gym, or

with friends [49]. These concepts are very similar to the features described in this paper for app use, but extend the information captured to include data reflecting not only the tools designed to promote behavior change, but the behavior change itself.

Conclusions

In an era where smartphone use is high and their presence ubiquitous, the ability to compare DMH data across different app suites and behavioral strategies for tracking mental well-being promises significant leverage and substantial impact on digital mental health care.

In conclusion, we have shown that five features defining the patterns of multiple digital mental health apps interaction show reproducibility across different deployment contexts and an increased rationale relative to currently used engagement metrics. Those five features provide information on the basic use of apps with different underlying behavioral strategies, a measure of the user's distribution of relative attention to these apps, the users' circadian use, their consecutive commitment to a specific app strategy, and users' interaction trajectory expressed by early engagement. Overall, we provide detailed, replicable concepts that allow for deeper insights into user engagement and comparability between different interventions as called for by research, e.g. Sieverink et al. [2].

As the field of digital mental health moves away from a single app, to considering how to optimize people's use of multiple apps in a platform, features such as these are both expected to improve AI performance as well as explanatory power and can be used as a starting point to conduct interpretable sensitivity analyses [3]. Therefore, this work lays the groundwork for measuring and understanding app use at a more granular level, that will in turn inform the design, tailoring, and delivery of apps to maximize user engagement and symptomatic improvement.

Abbreviations

eHealth: Healthcare practice supported by electronic processes and communication; DMH: Digital mental health; DSMB: Data safety monitoring board; GCP: Good clinical practice; GPS: Global positioning system; HIPAA: Health insurance portability and accountability act; IQR: Interquartile range; IRB: institutional review board; PD: Public deployment; RT: Research trial; $\bar{x}$: Average of a sample; $\hat{x}$: Median of a sample;

Declarations

Ethics approval and consent to participate

IntelliCare research trials were subject to ethics processes that have been approved by Northwestern University institutional review board (IRB) and were supervised by a specific data safety monitoring board (DSMB). Related activities with ethics processes that precede the 2013 Helsinki declaration, are ethical to the best of the authors' knowledge. The IRB and DSMB assured human subject protection, HIPAA, and GCP compliance adhering to commonly agreed standards set by international medical organizations.

Details on the IntelliCare project and primary outcomes may be found in previous reports [27, 28]. Data from the public deployment was acquired using similar measures and materials (apps) after users acknowledged that their deidentified data would be used for research purposes.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Funding

We acknowledge support by the Open Access Publication Fund of TU Berlin.

Author's contributions

All authors contributed in writing this article. MMT proposed the main ideas, wrote most of the first draft and the paper, and performed data analysis. AK and JN helped with background and objective scoping. AK, JN, and DCM helped with positioning. RZ contributed to the conceptualization, some interpretations and the conclusion. DCM supervised every step and assisted with some hypothesis generation, conventional measures, and provided critical review and valuable input. All authors read and approved the final manuscript.

Availability of data and materials

The clinical data that support the findings of this study cannot be shared at this time as they also indicate parts of ongoing studies, are bound to purpose as reported in CinicalTrials.gov (NCT02176226 and NCT02801877), and restrictions apply according to human subject protections. Additionally, under the consent given for the public deployment access is limited to the team of researchers and software developers at the Center for Behavioral Intervention Technologies (CBITs) at Northwestern University, and so not publicly available.

The IntelliCare suite used in this paper has been created by the CBITs at Northwestern University, Feinberg School of Medicine and is available to download for free in the Google Play Store at https://play.google.com/store/apps/developer?id=CBITs. Further descriptions and information about IntelliCare, its apps and components, can be found at https://intellicare.cbits.northwestern.edu/.

Acknowledgments

Not applicable.

Author details

[1]Technische Universität Berlin, Chair of Information and Communication Management, Strasse des 17. Juni 135, 10623 Berlin, Germany. [2]Center for Behavioral Intervention Technologies, Department of Preventive Medicine, Northwestern University, 750 N Lake Shore Dr, 60611 Chicago, United States. [3]Orygen, 35 Poplar Rd, 3052 Melbourne, Australia. [4]The Centre for Youth Mental Health, The University of Melbourne, 35 Poplar Rd, 3052 Melbourne, Australia.

References

1. Peterson, E.T., Carrabis, J.: Measuring the immeasurable: Visitor engagement. Web Analytics Demystified **14**, 16 (2008)
2. Sieverink, F., Kelders, S.M., Gemert-Pijnen, J.E.v.: Clarifying the Concept of Adherence to eHealth Technology: Systematic Review on When Usage Becomes Adherence. Journal of Medical Internet Research **19**(12), 402 (2017). doi:10.2196/jmir.8578. Accessed 2020-01-29
3. Holzinger, A., Dehmer, M., Jurisica, I.: Knowledge discovery and interactive data mining in bioinformatics-state-of-the-art, future challenges and research directions. BMC bioinformatics **15**(6), 1 (2014). doi:10.1186/1471-2105-15-S6-I1
4. Sandstrom, G.M., Lathia, N., Mascolo, C., Rentfrow, P.J.: Opportunities for Smartphones in Clinical Care: The Future of Mobile Mood Monitoring. The Journal of Clinical Psychiatry **77**(2), 135–137 (2016). doi:10.4088/JCP.15com10054. Accessed 2019-04-22
5. Feurer, M., Klein, A., Eggensperger, K., Springenberg, J., Blum, M., Hutter, F.: Efficient and Robust Automated Machine Learning. In: Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M., Garnett, R. (eds.) Advances in Neural Information Processing Systems 28, pp. 2962–2970. Curran Associates, Inc., New York (2015). http://papers.nips.cc/paper/5872-efficient-and-robust-automated-machine-learning.pdf Accessed 2019-08-20
6. Yarkoni, T., Westfall, J.: Choosing prediction over explanation in psychology: Lessons from machine learning. Perspectives on Psychological Science **12**(6), 1100–1122 (2017). doi:10.1177/1745691617693393. PMID: 28841086
7. DeMasi, O., Kording, K., Recht, B.: Meaningless comparisons lead to false optimism in medical machine learning. PLOS ONE **12**(9), 0184604 (2017). doi:10.1371/journal.pone.0184604. arXiv: 1707.06289. Accessed 2019-08-20
8. Saeb, S., Lonini, L., Jayaraman, A., Mohr, D.C., Kording, K.P.: Voodoo Machine Learning for Clinical Predictions. bioRxiv, 059774 (2016). doi:10.1101/059774. Accessed 2019-08-20
9. Ioannidis, J.P.A.: Why Most Published Research Findings Are False. PLOS Medicine **2**(8), 124 (2005). doi:10.1371/journal.pmed.0020124. Accessed 2019-04-22
10. Coiera, E., Ammenwerth, E., Georgiou, A., Magrabi, F.: Does health informatics have a replication crisis? Journal of the American Medical Informatics Association **25**(8), 963–968 (2018). doi:10.1093/jamia/ocy028. Accessed 2019-08-19
11. Hutson, M.: Artificial intelligence faces reproducibility crisis. Science **359**(6377), 725–726 (2018). doi:10.1126/science.359.6377.725. Accessed 2019-08-19
12. Canzian, L., Musolesi, M.: Trajectories of Depression: Unobtrusive Monitoring of Depressive States by Means of Smartphone Mobility Traces Analysis. In: Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing. UbiComp '15, pp. 1293–1304. ACM, New York, NY, USA (2015). doi:10.1145/2750858.2805845. event-place: Osaka, Japan
13. Harari, G.M., Müller, S.R., Aung, M.S., Rentfrow, P.J.: Smartphone sensing methods for studying behavior in everyday life. Current Opinion in Behavioral Sciences **18**, 83–90 (2017). doi:10.1016/j.cobeha.2017.07.018. Accessed 2019-08-20
14. Mohr, D.C., Zhang, M., Schueller, S.M.: Personal Sensing: Understanding Mental Health Using Ubiquitous Sensors and Machine Learning. Annual Review of Clinical Psychology **13**(1), 23–47 (2017). doi:10.1146/annurev-clinpsy-032816-044949

15. Asselbergs, J., Ruwaard, J., Ejdys, M., Schrader, N., Sijbrandij, M., Riper, H.: Mobile Phone-Based Unobtrusive Ecological Momentary Assessment of Day-to-Day Mood: An Explorative Study. Journal of Medical Internet Research **18**(3), 72 (2016). doi:10.2196/jmir.5505
16. Fleming, T., Bavin, L., Lucassen, M., Stasiak, K., Hopkins, S., Merry, S.: Beyond the Trial: Systematic Review of Real-World Uptake and Engagement With Digital Self-Help Interventions for Depression, Low Mood, or Anxiety. Journal of Medical Internet Research **20**(6), 199 (2018). doi:10.2196/jmir.9275. Accessed 2019-10-18
17. Pham, Q., Graham, G., Carrion, C., Morita, P.P., Seto, E., Stinson, J.N., Cafazzo, J.A.: A Library of Analytic Indicators to Evaluate Effective Engagement with Consumer mHealth Apps for Chronic Conditions: Scoping Review. JMIR mHealth and uHealth **7**(1), 11941 (2019). doi:10.2196/11941. Accessed 2020-01-29
18. Anderson, M.R., Antenucci, D., Bittorf, V., Burgess, M., Cafarella, M.J., Kumar, A., Niu, F., Park, Y., Ré, C., Zhang, C.: Brainwash: A data system for feature engineering. In: Cidr (2013)
19. Torous, J., Wisniewski, H., Liu, G., Keshavan, M.: Mental Health Mobile Phone App Usage, Concerns, and Benefits Among Psychiatric Outpatients: Comparative Survey Study. JMIR Mental Health **5**(4), 11715 (2018). doi:10.2196/11715. Accessed 2019-10-18
20. Collaboration, O.S.: Estimating the reproducibility of psychological science. Science **349**(6251), 4716 (2015). doi:10.1126/science.aac4716. Accessed 2019-04-22
21. Saeb, S., Lonini, L., Jayaraman, A., Mohr, D.C., Kording, K.P.: The need to approximate the use-case in clinical machine learning. GigaScience **6**(5), 1–9 (2017). doi:10.1093/gigascience/gix019. Accessed 2019-04-22
22. Marinescu, I.E., Lawlor, P.N., Kording, K.P.: Quasi-experimental causality in neuroscience and behavioural research. Nature Human Behaviour **2**(12), 891 (2018). doi:10.1038/s41562-018-0466-5. Accessed 2019-04-22
23. Taki, S., Lymer, S., Russell, C.G., Campbell, K., Laws, R., Ong, K.-L., Elliott, R., Denney-Wilson, E.: Assessing User Engagement of an mHealth Intervention: Development and Implementation of the Growing Healthy App Engagement Index. JMIR mHealth and uHealth **5**(6), 89 (2017). doi:10.2196/mhealth.7236. Accessed 2020-01-29
24. Ritterband, L.M., Thorndike, F.P., Cox, D.J., Kovatchev, B.P., Gonder-Frederick, L.A.: A Behavior Change Model for Internet Interventions. Annals of Behavioral Medicine **38**(1), 18–27 (2009). doi:10.1007/s12160-009-9133-4. Accessed 2019-10-17
25. Mohr, D.C., Weingardt, K.R., Reddy, M., Schueller, S.M.: Three Problems With Current Digital Mental Health Research . . . and Three Things We Can Do About Them. Psychiatric Services **68**(5), 427–429 (2017). doi:10.1176/appi.ps.201600541. Accessed 2019-10-18
26. Baumel, A., Kane, J.M.: Examining Predictors of Real-World User Engagement with Self-Guided eHealth Interventions: Analysis of Mobile Apps and Websites Using a Novel Dataset. Journal of Medical Internet Research **20**(12), 11491 (2018). doi:10.2196/11491. Accessed 2019-10-18
27. Mohr, D.C., Tomasino, K.N., Lattie, E.G., Palac, H.L., Kwasny, M.J., Weingardt, K., Karr, C.J., Kaiser, S.M., Rossom, R.C., Bardsley, L.R., Caccamo, L., Stiles-Shields, C., Schueller, S.M.: IntelliCare: An Eclectic, Skills-Based App Suite for the Treatment of Depression and Anxiety. Journal of Medical Internet Research **19**(1), 10 (2017). doi:10.2196/jmir.6645. Accessed 2019-08-15
28. Mohr, D., Schueller, S., Tomasino, K., M. Kaiser, S., Alam, N., Karr, C., Vergara, J., G. Gray, E., Kwasny, M., Lattie, E.: Randomized Trial Comparing the Effects of Coaching and Receipt of App Recommendations on Depression, Anxiety, and App Use in the IntelliCare Platform (Preprint). Journal of Medical Internet Research (2019). doi:10.2196/13609
29. Lattie, E.G., Schueller, S.M., Sargent, E., Stiles-Shields, C., Tomasino, K.N., Corden, M.E., Begale, M., Karr, C.J., Mohr, D.C.: Uptake and usage of IntelliCare: A publicly available suite of mental health and well-being apps. Internet Interventions **4**, 152–158 (2016). doi:10.1016/j.invent.2016.06.003. Accessed 2019-08-15
30. Stiles-Shields, C., Montague, E., Mohr, D.C.: The use of scenario-based design for the development of behavioral intervention technologies., Washington, DC (2016)
31. Kwasny, M.J., Schueller, S.M., Lattie, E., Gray, E.L., Mohr, D.C.: Exploring the Use of Multiple Mental Health Apps Within a Platform: Secondary Analysis of the IntelliCare Field Trial. JMIR Mental Health **6**(3), 11572 (2019). doi:10.2196/11572. Accessed 2020-01-27
32. Mohr, D.C., Schueller, S.M., Tomasino, K.N., Kaiser, S.M., Alam, N., Karr, C., Vergara, J.L., Gray, E.L., Kwasny, M.J., Lattie, E.G.: Comparison of the Effects of Coaching and Receipt of App Recommendations on Depression, Anxiety, and Engagement in the IntelliCare Platform: Factorial Randomized Controlled Trial. Journal of Medical Internet Research **21**(8), 13609 (2019). doi:10.2196/13609. Accessed 2020-01-27
33. Baumel, A., Muench, F., Edan, S., Kane, J.M.: Objective User Engagement With Mental Health Apps: Systematic Search and Panel-Based Usage Analysis. J Med Internet Res **21**(9), 14567 (2019). doi:10.2196/14567
34. Kelders, S.M., Kok, R.N., Ossebaard, H.C., Van Gemert-Pijnen, J.E.: Persuasive system design does matter: A systematic review of adherence to web-based interventions. J Med Internet Res **14**(6), 152 (2012). doi:10.2196/jmir.2104
35. Hofmann, H., Wickham, H., Kafadar, K.: Letter-Value Plots: Boxplots for Large Data. Journal of Computational and Graphical Statistics **26**(3), 469–477 (2017). doi:10.1080/10618600.2017.1305277
36. Rubin, G.: Better Than Before: Mastering the Habits of Our Everyday Lives. Broadway Books, Crown, Hachette UK (2015). Google-Books-ID: XbdvBQAAQBAJ
37. Murtagh, A.M., Todd, S.A.: Self-regulation: A challenge to the strength model. Journal of Articles in Support of the Null Hypothesis **3**(1), 19–51 (2004)
38. Stosny, S.: Self-Regulation (2011). http://www.psychologytoday.com/blog/anger-in-the-age-entitlement/201110/self-regulation Accessed 2019-10-16
39. Germain, A., Kupfer, D.J.: Circadian rhythm disturbances in depression. Human Psychopharmacology: Clinical and Experimental **23**(7), 571–585 (2008). doi:10.1002/hup.964. Accessed 2019-10-16
40. Phillips, L.A., Leventhal, H., Leventhal, E.A.: Assessing theoretical predictors of long-term medication adherence: Patients' treatment-related beliefs, experiential feedback and habit development. Psychology &

Health **28**(10), 1135–1151 (2013). doi:10.1080/08870446.2013.793798. PMID: 23627524
41. Schueller, S.M., Mohr, D.C.: Initial field trial of a coach-supported web-based depression treatment. In: 2015 9th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth), pp. 25–28 (2015). IEEE
42. Neal, D.T., Wood, W., Quinn, J.M.: Habits—A repeat performance. Current Directions in Psychological Science **15**(4), 198–202 (2006)
43. Ouellette, J.A., Wood, W.: Habit and intention in everyday life: The multiple processes by which past behavior predicts future behavior. Psychological bulletin **124**(1), 54 (1998)
44. Webb, T., Joseph, J., Yardley, L., Michie, S.: Using the Internet to Promote Health Behavior Change: A Systematic Review and Meta-analysis of the Impact of Theoretical Basis, Use of Behavior Change Techniques, and Mode of Delivery on Efficacy. Journal of Medical Internet Research **12**(1), 4 (2010). doi:10.2196/jmir.1376. Accessed 2019-10-17
45. Huang, H.-Y., Bashir, M.: Users' Adoption of Mental Health Apps: Examining the Impact of Information Cues. JMIR mHealth and uHealth **5**(6), 83 (2017). doi:10.2196/mhealth.6827. Accessed 2019-10-17
46. Larsen, M.E., Nicholas, J., Christensen, H.: Quantifying app store dynamics: Longitudinal tracking of mental health apps. JMIR Mhealth Uhealth **4**(3), 96 (2016). doi:10.2196/mhealth.6020
47. Saeb, S., Zhang, M., Karr, C.J., Schueller, S.M., Corden, M.E., Kording, K.P., Mohr, D.C.: Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: An exploratory study. J Med Internet Res **17**(7), 175 (2015). doi:10.2196/jmir.4273
48. Saeb, S., Lattie, E.G., Schueller, S.M., Kording, K.P., Mohr, D.C.: The relationship between mobile phone location sensor data and depressive symptom severity. PeerJ **4**, 2537 (2016)
49. Saeb, S., Lattie, E.G., Kording, K.P., Mohr, D.C.: Mobile phone detection of semantic location and its relationship to depression and anxiety. JMIR Mhealth Uhealth **5**(8), 112 (2017). doi:10.2196/mhealth.7297

Figures

figures/fig_usage.pdf

Figure 1 Usage visualization. On a global level (left), each app converges to its inherent average level of expected events by time, given by design. For a given app (App 2), the user's average session pace and specific recent deviations can be accounted for on an individual level (right).

figures/fig_attention.pdf

Figure 2 Allocated Attention visualization. The most used app (App 1) is the reference to score all other downloaded apps relative share, where abandoned apps only have minor influence (App 4 and App 5).

figures/fig_circadian.pdf

Figure 3 Circadian Use visualization. Starting from any given session today (right), a fixed window centered around the same timestamp yesterday is defined (center). The root mean square of deviations within that fixed window is a proxy for *Circadian Use*, and can be repeated similarly for more previous days (left).

figures/fig_commitment.pdf

Figure 4 Follow-Up Commitment visualization. In the upper row of bend arrows, App 1 is used twice in a consecutive manner (on the very left and very right side), while it is used five times overall. Longer, more complex sequences of use are omitted for simplification (center).

figures/fig_trajectory.pdf

Figure 5 App Use Trajectory visualization. After the initial download, a phase of low interaction density occurs (1st Day), followed by a high density of sessions (2nd Day) where the app is explored. The major phase (2nd Week) shows regular sessions, before a fade out of use occurs in the end (right).

Tables

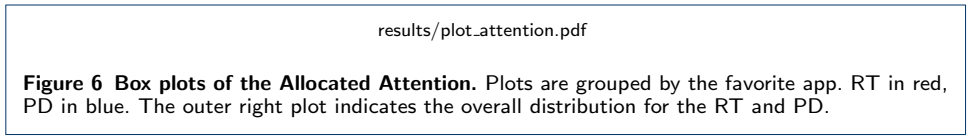

Figure 6 Box plots of the Allocated Attention. Plots are grouped by the favorite app. RT in red, PD in blue. The outer right plot indicates the overall distribution for the RT and PD.

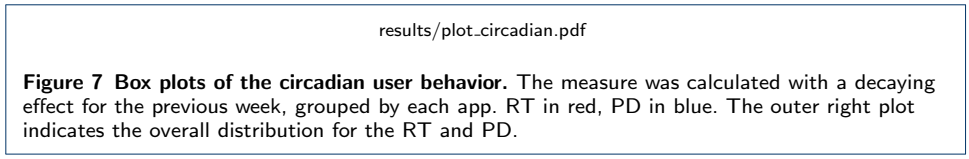

Figure 7 Box plots of the circadian user behavior. The measure was calculated with a decaying effect for the previous week, grouped by each app. RT in red, PD in blue. The outer right plot indicates the overall distribution for the RT and PD.

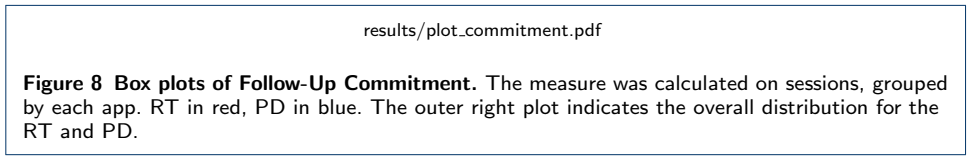

Figure 8 Box plots of Follow-Up Commitment. The measure was calculated on sessions, grouped by each app. RT in red, PD in blue. The outer right plot indicates the overall distribution for the RT and PD.

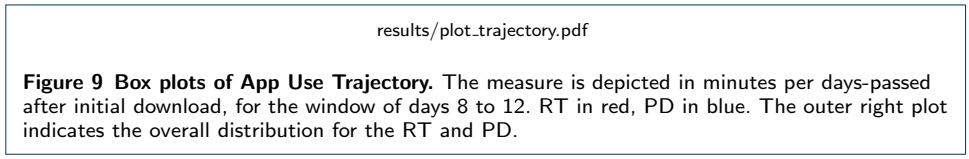

Figure 9 Box plots of App Use Trajectory. The measure is depicted in minutes per days-passed after initial download, for the window of days 8 to 12. RT in red, PD in blue. The outer right plot indicates the overall distribution for the RT and PD.

Table 1 Correlation matrix for Adjusted Usage. RT lower half, PD upper half. A 2 tailed p-value < 0.05 is flagged with one star (*), a p-value < 0.01 is flagged with two stars (**), and all other p-values, that are < 0.001, are not flagged for readability purpose.

		Public deployment correlations						
		Usage	Sessions	Events	Time	Freq.	Lifetime	Apps
Research trial correlations	Usage		0.89	0.94	0.99	0.19	0.23	0.15
	Sessions	0.78		0.97	0.91	0.18	0.17	0.11
	Events	0.91	0.84		0.95	0.23	0.20	0.15
	Time	0.98	0.73	0.91		0.19	0.22	0.14
	Freq.	0.63	0.42	0.68	0.65		−0.34	0.31
	Lifetime	0.26	0.29	0.21	0.25	−0.32		0.04*
	Apps	0.31	0.28	0.28	0.31	0.19**	0.15	