User Engagement Metrics and Patterns in Phendo, an Endometriosis Research Mobile App

Noemie Elhadad (✉ noemie.elhadad@columbia.edu)  
Columbia University  https://orcid.org/0000-0001-9721-5240

Iñigo Urteaga  
Columbia University  https://orcid.org/0000-0003-3656-0037

Sharon Lipsky-Gorman  
Columbia University

Mollie McKillop  
IBM Watson Health

Article

Keywords:

Posted Date: July 20th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1862924/v1

License: ☑️ This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License
User Engagement Metrics and Patterns in Phendo, an Endometriosis Research Mobile App

Iñigo Urteaga†a,b, Sharon Lipsky-Gorman†c, Mollie McKillopd, and Noémie Elhadad*b,c

† Equal contribution

a Department of Applied Physics and Applied Mathematics, Columbia University, New York NY 10027
b Data Science Institute, Columbia University, New York NY 10027
c Department of Biomedical Informatics, Columbia University, New York NY 10032
d IBM Watson Health, New York, NY, USA (Work done while at the Department of Biomedical Informatics, Columbia University)

* Corresponding author, noemie.elhadad@columbia.edu

July 15, 2022
Abstract

We characterize short-term and long-term user engagement patterns in a self-tracking, mobile health app. We introduce and define engagement metrics to capture the quantity, duration, and density of participant engagement according to different domains of self-tracking. We focus our study on Phendo, a research app designed for participants to self-track their experiences of endometriosis—a chronic disease in women of reproductive age. Given a cohort of Phendo participants with at least one initial week of use, we analyze their engagement patterns based on amount and timing of daily app usage and stratify them according to their short-term engagement patterns, i.e., within their first 12 weeks of app-use. Given the overall cohort and its stratified short-term participant engagement groups, we then assess overall longitudinal engagement patterns with the app beyond the first 12 weeks, as well as with disease-specific self-tracking domain types. We identify four groups of participants in the Phendo cohort (n = 4,993) according to their short-term engagement: Regulars, Usuals, Occasionals, and Seldoms. Participants across the groups do not differ in disease status or demographics, except for age and education. We find that, while the stratification is based on a short-term period, the groups continue to exhibit distinct longitudinal engagement patterns in the long-term (up to 4 years). We also find that Regulars are more likely to engage with self-tracking related to disease self-management than the other groups. These findings have implications for the design of research mobile apps, as certain functionalities like tracking of self-management, might yield longer and richer engagement with certain types of participants, even if self-management is not part of the original intent of the research app. More broadly, our proposed engagement metrics and analyses provide a roadmap for exploring participant engagement patterns in mobile research apps.

Introduction

The widespread use of smartphones across all segments of population and the development of mobile health (mHealth) technology has made possible new ways of supporting patients to manage their disease, and for healthcare providers to connect with their patients [1]. For researchers, smartphone apps represent an exciting opportunity to engage directly with a very large number of participants and to reach populations that are typically under-studied [2]. Furthermore, through research apps, participants can contribute new types of data and at a finer temporal resolution than in traditional research settings [1][3]. Research apps are thus becoming a critical instrument in observational studies for disease characterization: through daily reports from participants, researchers can get new insights into diseases and capture temporal fluctuations that might have escaped traditional research instruments [4].

Research apps have been designed for a number of conditions [5][11]. mHealth research studies have been successful in achieving high app download rates and in enrolling very large numbers of participants, but retention of participants is still a challenge. These apps often see drastic decreases in long-term app usage, e.g., up to 99% decrease in engagement over 6 months [6][10][12]. This is a missed opportunity, as longer retention may bring additional insights into physiological phenomena and diseases, as their presentation varies through time [13]. Longer retention may also facilitate more rigorous data interpretation, for instance, by assessing and mitigating for data missing not at random [14][10].

Previous work has investigated indicators of participant retention in research apps that can inform participant recruitment [2]. Specifically, they explored patterns of engagement in research apps in the first 12 weeks of use, and identified that four groups of participants emerge across all research apps investigated [2]. In this study, we carry out a similar analysis, yet investigate patterns of engagement with a research app both in the short- and long-term.

To that end, we present engagement metrics that capture how much, how often, for how long, and with which types of questions participants engage. Elucidating who the highly engaged participants are and their specific patterns of engagement can inform recruitment strategies, as well as which app functionalities to design to promote long-term engagement in a research study.

The context for this work is an observational study to characterize and phenotype endometriosis, carried out through Phendo [17][18], a mobile research app designed for participants to self-track their experiences of endometriosis and contribute to its phenotyping. In the following, we characterize short-term and long-term user engagement patterns in Phendo. We introduce and define engagement metrics to capture the quantity, duration, and density of participant engagement according to different domains of self-tracking. We analyze Phendo user engagement patterns based on amount and timing of daily app usage, and stratify them according to their short-term engagement patterns (first 12 weeks of use). Given the overall cohort and its stratified participant short-term engagement groups, we then assess overall longitudinal engagement patterns with the app beyond the initial 12 weeks, i.e., from onboarding to their last app interaction within the study period, as well as with disease-specific self-tracking domain types.
Results

Phendo Cohort. Starting from \( N = 15,000 \) total Phendo users who joined between November 2016 and April 2021, 12,466 participants reported an endometriosis diagnosis. Among these participants, 551 (4.4%) opted out from the study. From the remaining 11,915 participants, we excluded those who self-tracked for less than a 7-day span (7,475 participants, 62%). The resulting Phendo cohort comprises \( N = 4,993 \) participants. Table 1 summarizes the participants’ demographics and disease-related information of the Phendo cohort, as extracted from their Phendo profile and WERF information described in the Methods sections.

Participants in the Phendo cohort engaged with all Phendo self-tracking questions (described in the Methods sections and Table 3), for a total of 820,173 observations across the 16 self-tracking questions, with counts per Phendo questions depicted in Figure 1. The most self-tracked questions relate to overall quality of day (“How was your day?”) and menstrual cycle tracking (“Do you have your period?”). The least self-tracked questions relate to the presence of bleeding outside of period (“If you experienced bleeding, what type?”) and sexual activity (“Did you have sex and if so how was it?”). The low frequency of the sexual activity question can be in part explained because Phendo does not include this question in its default configuration.

![Figure 1: Daily engagement counts for all self-tracking questions across the Phendo cohort.](image)

Short-Term Engagement Analysis. The short-term, daily app-usage analysis, based on the first 12 weeks of use of participants, yielded 4 groups, which we refer to as Regulars (\( N = 339 \)), Usuals (\( N = 769 \)), Occasionals (\( N = 1,459 \)), and Seldoms (\( N = 2,426 \)). Table 1 summarizes the participants’ demographics and disease-related information, as extracted from their Phendo profile and WERF information, for each engagement group.

The engagement groups are all similar across demographics (race/ethnicity, gender, height, weight, employment, relationship status, living environments), except in age (\( p < 10^{-5} \)) and education (\( p < 10^{-5} \)) —the full statistical analysis is provided in the Supplementary Information. There are no significant differences across the participant groups with respect to disease-specific variables, such as number of years since endometriosis diagnosis, number of years experiencing symptoms, and number of laparoscopies undergone (laparoscopic surgeries are both diagnostic and treatment procedures for endometriosis).

Figure 2 illustrates per-day engagement vectors over the initial 12 week of app use for the Phendo cohort. Daily engagement patterns vary across each stratification group, both in density and in time. Regulars have the greatest daily engagement with Phendo: their activity is densest and covers the full 12-week interval. Usuals engage in self-tracking quite frequently, but not as densely as Regulars. Occasionals engage not only less frequently, but also for a shorter span. Seldoms are at the other extreme of the daily engagement spectrum, as they interact at the very beginning, but seldomly return to use the app.

Long-Term Engagement Analysis. Figure 3 shows a histogram of the long-term engagement metrics for the Phendo cohort, and Table 2 provides summary statistics for each engagement group.

For the entire Phendo cohort, the median day-span is 42 days and the median number of distinct days tracked is 8. For Regulars, Usuals, Occasionals, and Seldoms the median day-span are 183 days, 94 days, 32 days, and 24 days
respectively; and the median number of distinct days tracked are 91 days, 31 days, 11 days, and 4 days respectively. We observe that both histograms have a long tail, with some highly engaged participants who regularly self-track for 4+ years.

Even though participant stratification is based on short-time (initial 12 weeks) engagement, these patterns correlate with participants’ entire longitudinal engagement of up to 4 years. This dependency is visible both according to the day-span metric (Figure 3a; e.g., the median day-span for Seldoms of 24 days vs. 183 days for Regulars) and the longitudinal distinct number of days metric (Figure 3b; e.g., Seldoms track a median of only 4 distinct days vs. 91 days for Regulars).

(a) Number of day-spans over which a participant tracked. (b) Number of distinct days tracked by a participant.

Figure 3: Longitudinal engagement metrics: (a) number of participants per day-spans for which they engaged with Phendo, which indicates the duration between their first self-tracking to their last self-tracking activity at the time of analysis; and (b) number of participants per distinct days, which indicates the individual number of days a participant engaged in self-tracking activity on the app. In (a) the red circles indicate the median span for each engagement group; in (b) the red circle indicates the median days tracked for each engagement group.

Domain and Question-based Engagement Patterns. We analyze whether different engagement groups exhibit different self-tracking patterns, according to the domains and questions they engage with. Figure 4 shows the domain-engagement ratios for the Phendo cohort across each engagement group, along with their mean and standard deviation.
Figure 4: Domain engagement ratios for each engagement group. The domains are ordered by statistical difference; ** indicates statistically significant results with \( p < 0.001 \).

For the three domains (“How was my day,” “How I self-manage,” and “What happens to me”) described in Table 3, we observe significantly different \( p < 10^{-5} \) domain engagement ratios across the groups, indicating heterogeneous engagement behaviors across groups (statistical analysis results are provided in the Supplementary Information).

Longitudinal engagement ratios for Regulars are statistically different for the self-management and day-related domains \( p < 10^{-4} \) for all pairwise, Bonferroni corrected, Kolmogorov-Smirnov and Welch t-tests), but not for the “What happens to me” domain. Based on these results, and by inspecting each ratio’s per-group probability density functions (Figure S3 in Supplementary Information), we conclude that Regulars engage more than other groups with both the “How I self-manage” and “How is my day” domains.

To further elucidate how engagement within a domain differs across participant groups, Figure 5 shows engagement ratios at the question level within the self-management domain. We find that Regulars’ engagement with the supplements question is significantly different from other groups (pairwise, Bonferroni corrected, Kolmogorov-Smirnov, Mann-Whitney and Welch’s t-tests with \( p < 10^{-4} \)). The exercise, food and hormone question engagement ratios also appear to be slightly different for users within the Regulars group. In contrast, the ratios for the remaining questions within the “How I self-manage” domain are not significantly different across engagement stratifications.

Figure 5: Question engagement ratios for the self-management domain across the engagement groups. The questions are ordered by statistical difference, ** indicates statistically significant results with \( p < 0.001 \).
Discussion

In this study, we introduce several engagement metrics to accommodate complementary views of mHealth self-tracking behavior: how much (i.e., the distinct number of days), for how long (i.e., the span of days from first to last app-usage), and how often (question and domain engagement ratios) participants engage with a research app. These metrics are more granular than the traditional mobile app retention and churn rates, and aim to inform which types of functionalities (in our case self-tracking domains) are associated with continued engagement for capturing meaningful disease-specific information.

The cohort for this study comprises Phendo participants who engaged in self-tracking over a span of 7 days or longer. Four distinct groups of participants (Regulars, Usuals, Occasionals, and Seldoms) emerged when accounting for participants’ daily app-usage in their first 12 weeks after onboarding. While these groups differ in their engagement patterns with the app over their first three months, there are a number of insights that arise from our analyses.

First, participant across groups do not differ with respect to their endometriosis diagnosis or status, suggesting that engagement with the research app is independent from the health status of participants. This confirms earlier findings, where no correlation was found between severity of participant’s disease and their recorded number of self-tracked days or of self-tracking events (19). Disease status does play a significant role in onboarding into the app, and thus recruitment strategies should focus on individuals across disease stages for variability. But once onboarded, we find app engagement not to be dependent on disease status.

Second, the patterns of engagement for the first 12 weeks are correlated with longitudinal engagement patterns. Participants who engage with Phendo consistently (in density and duration) within the initial 12 weeks show sustained, higher engagement (both in terms of distinct number of days and longer day-spans) beyond the first 12 weeks of app use. This finding has interesting implications for monitoring engagement through time. If the first three months after onboarding play an important role towards continued engagement (20), engagement interventions should occur during that time period rather than later. This also raises mHealth design questions: are there app design choices that capture certain participants’ interest better than others? An open research avenue is to investigate whether adding/removing these functionalities is an effective (disease specific) mHealth engagement technique.

Third, the participant groups show different patterns of engagement with regards to the type of questions they self-track. In particular, Regular participants were more likely to answer questions related to their disease self-management (specifically regarding exercise, food, and sex) than other groups. These findings reflect a motivated Phendo user, i.e., one determined to document their daily endometriosis management towards a better and more nuanced understanding of their disease experience. Even though research app participants are recruited to contribute data to understanding a disease, functionalities that support self-management (and thus go beyond simply collecting data) might be helpful to ensure sustained engagement. We note, however, that our findings are based on correlations and not causal models, and thus we are cautious in our claims.

There are likely many factors explaining why different users engage differently with Phendo (e.g., why do Regulars engage more with self-management questions). For instance, Autonomy, a key concept of self-determination theory, is a motivator for the self-management of daily symptoms of endometriosis (21), and therefore, self-determination theory may explain participants’ internal motivation to self-manage and the observed engagement (22 23). Nevertheless, these findings suggest interesting directions for the design of research apps that include support for self-management as a core functionality, even if it is not part of the original goal of the research app. For instance, different sophisticated disease self-management features (e.g., recommendations for strategies, provision of insights to support reflection, or even interaction functionalities for checking on disease management with providers (21)) may engage different types of participants. Follow-up studies will be necessary to elucidate this hypothesis.

Limitations. There are a number of limitations with this study and the findings it suggests for engagement and research apps. The population of participants might not represent all endometriosis patients. In particular, our cohort has access to a smartphone and is technologically literate, and the median age (28) indicates a population with younger age at diagnosis compared to what is reported in the literature. Our analysis reveals correlations, but further investigation is needed to elucidate causality and the directionality of these associations. The study focuses on a single condition, and some of the engagement patterns identified in this work, while non-specific to endometriosis, might not generalize to other types of chronic conditions. We also note that our analysis is limited to an aggregated, fixed-time (either initial 12 weeks or full app-usage) view of engagement, and do not characterize each participant’s engagement within their daily or weekly Phendo app-usage timeline.
Conclusions. The engagement patterns in Phendo, a research app for self-tracking experience of endometriosis, are heterogeneous with respect to how often, how much, and for how long participants engage with the app, as well as in what disease-specific self-tracking domain (i.e., self-management) they are most engaged with. These insights suggest that participants’ motivation for self-management might be a critical factor for long-term engagement. Appropriately timed retention strategies and targeted app-functionality designs might incentivize overall engagement. Finally, we argue that understanding mHealth app engagement is a critical factor towards leveraging patient-generated data for generalizable medical research, and that the metrics we propose to quantify engagement from multiple aspects, can help inform who and how participants engage in mobile health research.

Methods

We first describe the Phendo research app and define the cohort of Phendo participants specific to this study. We then present the participant stratification method based on short-term daily app-usage, define longitudinal engagement metrics, and describe the statistical analyses to characterize long-term engagement patterns.

The Phendo Research App.

Phendo functionalities. Phendo is a research app that aims to characterize and phenotype endometriosis from patient’s experiences of disease. Endometriosis is a prevalent, chronic, and systemic inflammatory disorder in women of reproductive age (25). Participatory design was used to assess app feasibility, examine participants’ attitude towards self-tracking their disease, and identify what aspects of the disease to self-track (11, 26).

Through the design process, the following functionalities emerged: (1) ability to self-track a large number of disease dimensions to reflect the systemic impact of the disease; (2) ability to hide self-tracking questions that do not pertain to participants’ own experience of disease (e.g., menstruation or sexual activity); (3) ability to self-track some of the disease dimensions multiple times throughout a day (e.g., pain episodes in specific body locations); (4) ability to customize answers to some disease dimensions (e.g., medications); (5) ability to edit answers in case of erroneous entry; (6) ability to self-track retroactively; and (7) ability to access and review self-tracked answers through time.

Figure 6 shows example screenshots of Phendo, which was developed for iOS and Android phones and launched in November 2016 as a free research app on app stores. Data collected through the Phendo app is stored on a HIPAA-certified server. In addition to the self-tracked data, the database has an audit-trail which maintains a chronological log of participant tracking events.

Recruitment for Phendo was carried out through endometriosis advocacy groups. The research study is open to any individual 13 years and older (with assent for minors and consent from a guardian) who has reached menarche. Once a Phendo participant consents to the study, they can track as much or as little as they want in the app. Participants are suggested to stay engaged with the study for at least three months and longer if they desire to, and they have the option to either set a time for daily reminder notifications or opt out of notifications. Participants can opt-out of the study anytime through the app.
Phendo self-tracking questions and domains. Table provides the full list of self-tracking questions available in the app. They relate to pain, menstruation, bleeding, gastrointestinal (GI) and genitourinary symptoms, other symptoms such as fatigue, moods, quality of life, activities of daily living, exercise, foods, medications, supplements, hormones, self-management strategy, and sexual activity (which is hidden by default). For this study, we map the self-tracking questions into three domains: “How was my day”; “How I self-manage”, which comprises questions about self-management, as well as activities of daily living; and “What happens to me”, which includes questions about experienced symptoms.

Phendo profile and WERF data. At onboarding, participants are encouraged to fill out their Phendo profile, which collects basic demographic and disease-specific information, and an electronic version of the WERF EPHect survey—a standardized, detailed clinical questionnaire designed and validated for endometriosis research (27). As for the self-tracking questions, both the Phendo profile and WERF survey are visible to the participants, can be updated at any time, and can be filled in a voluntary fashion.

Phendo Cohort. The cohort considered for this engagement study comprises Phendo participants with the following inclusion criteria: (1) joined the study between November 2016 (launch of the app) and April 2021; (2) have not opted out from the study at the time of analysis (August 2021); (3) have an endometriosis diagnosis (whether reported as confirmed through laparoscopy, or not yet); and (4) have self-tracked over a span of at least 7 days. We censor the entry into the cohort from April to August 2021 because we want to observe engagement across 12 weeks at least.

For each participant in the cohort, we collect: (1) their longitudinal self-tracking data, as entered in Phendo from November 2016 to August 2021; (2) their Phendo profile information, and (3) their WERF survey responses. The self-tracked data is used for the analysis of engagement patterns, while the Phendo profile and WERF answers are analyzed to characterize learned participant groups.

Short-Term Engagement Analysis.

We carry out a short-term engagement analysis on the first 12 weeks of use of the Phendo cohort, in a similar fashion to previous work on exploring patterns of engagement in research apps (2). However, our methodology is novel in that we identify engagement clusters based on a lower dimensional subspace, via Bernoulli-distributed Principal Component Analysis, that captures the idiosyncrasies of the daily app-usage data.

We set the unit of engagement analysis to days, i.e., we consider a participant engaged with the app in a day if they self-tracked any question in the app once or multiple times that day. We keep track of this indicator for all self-tracking questions for each consecutive day of the first 12 weeks of each participant’s use of the app. Participants’ daily engagement data is represented as follows: for a given user, we compute a per-day engagement binary indicator (1 if the participant interacted with the app, 0 otherwise), and define a user’s engagement vector as the collection of these 84 engagement indicators for each consecutive day over their first 12 weeks. We use these per-participant, binary 84-dimensional vectors for their engagement stratification.

We propose to group or cluster these engagement vectors in a lower dimensional subspace (28). Because the engagement vectors are binary, we leverage advances in generalizing Principal Component Analysis (PCA) to non-Gaussian data and, in particular, to Bernoulli distributed data (29, 30). These generalized forms of PCA search for a low dimensional basis defining a surface that is close to all the observed data points. With Binary-PCA, we find a lower dimensional subspace in which the projected binary engagement vectors are “close” to each other based on an appropriate notion of distance. In Gaussian-PCA, this distance is computed based on the sum of the squared distances from the data points to their projections (i.e., via the Gaussian likelihood); while in Binary-PCA, it is based on the logistic regression link function —see (29) for a thorough mathematical definition and analysis of generalized PCA.

We select the basis-vector components that explain 80% of the observed variance in the learned Binary-PCA based subspace. In this real-valued lower-dimensional subspace, a notion of Euclidean distance allows us to identify different participant groups: i.e., we perform K-means clustering in this projected subspace (28). We use the elbow method to decide the number of groups that best stratify the data: i.e., we select the number of clusters that correspond to the point with maximum curvature of the sum of squared distances between each point and the centroid in a cluster (31). As output of this procedure executed on the participants’ initial 12-week binary engagement vectors, we obtain assignments of each participant to the learned stratification groups, based on the minimum distance of each projected engagement vector to the engagement cluster centroids.
We compare our results to alternative clustering approaches in the literature: k-means performed in the binary engagement vector space; k-means over standardized engagement vectors, i.e., after a z-score transformation that centers and scales the data to have mean 0 and standard deviation 1; and k-means in a subspace found via Gaussian-PCA on standardized engagement vectors. Results (provided in the Supplementary Information) showcase that participant stratification is robust to these alternative stratification approaches.

**Long-Term Engagement Metrics.**

To investigate patterns of engagement with the app in the long term, we propose engagement metrics that capture how much, how often, for how long and with which specific questions and domains a participant engages longitudinally.

We study users’ longitudinal engagement via the number of distinct days tracked by a participant and the number of day-spans over which participants tracked, calculated as the number of days between when participants joined the study, up until the day they last tracked anything. With these, we capture users’ engagement behavior over their Phendo timeline, i.e., from onboarding to their last app interaction within the study period.

We study users’ engagement frequency with the app via engagement ratios, i.e., how often each user self-tracks specific questions and domains over their app-usage timeline. These are normalized quantities that account for the distinct number of days each participant has engaged with the app, which mitigates the effect of different participant engagement behaviors. Given an app user, we compute the following:

- **Question engagement ratio** ($\lambda_{u,q}$): the number of days question $q$ was answered by user $u$, over the total number of days any question was answered by user $u$.

- **Domain engagement ratio** ($\lambda_{u,d}$): the ratio of the number of days a domain $d$ question was answered by user $u$, over the total number of days any question was answered by user $u$.

We note that these engagement metrics and ratios are computed over the participants’ full Phendo usage timeline, and are not limited to their initial 12 weeks. We compute full cohort and stratified engagement results for these metrics via several statistics of interest (e.g., histograms or median values); e.g., $p(\lambda_{g,q})$ indicates the probability distribution of the question engagement ratio for all users assigned to a specific group $g$. We characterize the longitudinal engagement of Phendo participants with the analysis of these metrics, and evaluate the significance of the obtained results according to different statistical analysis.

**Long-Term Engagement Analysis.**

To statistically compare the distributions of the computed longitudinal engagement metrics (across the entire period of engagement with the app, beyond the first 12 weeks of app-usage) across groups, we use the Kruskal-Wallis H-test of independence, a non-parametric method that accommodates different sample sizes, without a Gaussian distribution assumption, for testing whether samples originate from the same distribution. The null hypothesis is that the medians of all engagement groups are equal, while the alternative hypothesis is that at least one of the medians of a group is different from at least one other group’s median. A significant Kruskal–Wallis test indicates that at least one group sample stochastically dominates some other one.

Once a significant difference is found across engagement groups, and because the test does not identify per-group dominance, subsequent pairwise tests are performed, for which the Bonferroni correction is applied:

- **Kolmogorov-Smirnov** tests to quantify whether the empirical distributions of the engagement metrics differ between two groups. This nonparametric test measures the distance between the empirical cumulative distribution functions (CDFs) of two samples, and is sensitive to differences in both location and shape of the CDFs of the two samples, allowing us to establish whether the engagement metric distributions differ.

- **Mann–Whitney** tests to quantify whether the distribution of one sample dominates the other, and more specifically, if the probability of an observation from group $A$ exceeding an observation from group $B$ is different than the probability of an observation from $B$ exceeding an observation from $A$. Because the underlying metric distributions of the two groups differ in shape and not only in location, we are not testing for the hypothesis of (un)-equal medians, but that these engagement distributions differ.

- **Welch’s t-test** to quantify whether the empirical average of the engagement metrics are different between two groups. This is a two-sample location test for determining whether two populations have equal means, which assumes the two population sample means are normally distributed with unequal population variances. Even though our engagement metric distributions are non-normal (see histograms in the Supplementary
Information), sample means in large studies can be compared with this test even when the normality assumption does not hold at the population level, as the Welch’s t-test is robust to non-Gaussian (e.g., severely skewed) data (36 38 39).

**Ethics.** Data collection and the analysis presented in this work was carried out under Research Protocol #AAAQ9812 approved by Columbia University IRB. We obtained signed informed consent from all participants in the study.

**Data Availability**

Please contact the authors to obtain access to a de-identified version of the data that supports the findings of this study through a data-use agreement.

**Code Availability**

Our code has been developed using open source tools in Python with common statistical libraries (e.g., NumPy, SciPy and Pandas). The code required for processing the data and producing the presented results will be made available in a public GitHub repository upon acceptance.

**Acknowledgements**

The authors are deeply grateful to all Phendo users whose de-identified data have been used for this study and acknowledge the many endometriosis advocacy groups that helped in recruiting them.

This work was supported by National Library of Medicine awards T15 LM007079 (MM) and R01 LM013043 (IU, NE).

**Author Contributions**

NE, SLG, and MM conceived the proposed research. SLG prepared the data; NE, SLG, and IU designed the engagement metrics and experiments, conducted the analysis and interpreted the results. NE, SLG, and IU wrote the manuscript, and MM reviewed it. All authors read and approved the manuscript.

**Competing Interests**

All authors declare that they have no competing interests.
References


[37] Welch, B. L. The Generalization of ‘Student’s’ Problem when Several Different Population Variances are Involved. *Biometrika* 34, 28–35 (1947).


Figure legends

**Figure 1:** Daily engagement counts for all self-tracking questions across the Phendo cohort.

**Figure 2:** Per-day engagement binary vectors (a colored square indicates that a participant self-tracked that day) over the first 12 weeks of app usage for the Phendo cohort, across the four engagement groups: Regular, Usuals, Occasionals, and Seldoms.

**Figure 3:** Longitudinal engagement metrics: (a) number of participants per day-spans for which they engaged with Phendo, which indicates the duration between their first self-tracking to their last self-tracking activity at the time of analysis; and (b) number of participants per distinct days, which indicates the individual number of days a participant engaged in self-tracking activity on the app. In (a) the red circles indicate the median span for each engagement group; in (b) the red circle indicates the median days tracked for each engagement group.

**Figure 4:** Domain engagement ratios for each engagement group. The domains are ordered by statistical difference; ** indicates statistically significant results with \( p < 0.001 \).

**Figure 5:** Question engagement ratios for the self-management domain across the engagement groups. The questions are ordered by statistical difference; ** indicates statistically significant results with \( p < 0.001 \).

**Figure 6:** Phendo screenshots. The two panels on the left show the different questions participants can answer. The two panels on the right show the self-tracking screen for specific questions about self-management strategies tried that day and an overall assessment of the day.
Tables
# Demographics of the Phendo cohort.

<table>
<thead>
<tr>
<th>Variable</th>
<th>All</th>
<th>Regulars</th>
<th>Usuals</th>
<th>Occasional</th>
<th>Seldom</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>N (%)</strong></td>
<td>4,993(100)</td>
<td>339(6.8)</td>
<td>769(15.4)</td>
<td>1,459(29.2)</td>
<td>2,426(48.6)</td>
</tr>
<tr>
<td><strong>Age (years)</strong></td>
<td>median 28 (1-62)</td>
<td>14 (2-62)</td>
<td>29 (1-56)</td>
<td>28 (14-56)</td>
<td>28 (13-55)</td>
</tr>
<tr>
<td><strong>Weight (lbs)</strong></td>
<td>median 145 (50-440)</td>
<td>143 (84-350)</td>
<td>414 (81-407)</td>
<td>170 (70-440)</td>
<td>145 (50-396)</td>
</tr>
<tr>
<td><strong>Height (in)</strong></td>
<td>median 65 (1-62)</td>
<td>65 (1-62)</td>
<td>65 (15-56)</td>
<td>65 (14-56)</td>
<td>65 (13-55)</td>
</tr>
<tr>
<td><strong>How diagnosed n(%)</strong></td>
<td>Laparoscopy 3136 (62.8)</td>
<td>242 (71.4)</td>
<td>504 (65.5)</td>
<td>901 (61.8)</td>
<td>1489 (61.4)</td>
</tr>
<tr>
<td><strong>Education N(%)</strong></td>
<td>College + 2902 (58.1)</td>
<td>231 (68.1)</td>
<td>488 (63.5)</td>
<td>851 (58.3)</td>
<td>1332 (54.9)</td>
</tr>
<tr>
<td><strong>Employment N(%)</strong></td>
<td>Employed 3113 (62.3)</td>
<td>231 (68.1)</td>
<td>462 (60.1)</td>
<td>924 (61.3)</td>
<td>1496 (61.7)</td>
</tr>
<tr>
<td><strong>Race/Ethnicity N (%)</strong></td>
<td>White, Non-Hispanic 4060 (81.3)</td>
<td>298 (87.9)</td>
<td>632 (82.2)</td>
<td>1199 (82.2)</td>
<td>1931 (79.6)</td>
</tr>
<tr>
<td><strong>Gender N (%)</strong></td>
<td>Female 4577 (97.7)</td>
<td>331 (97.6)</td>
<td>750 (97.5)</td>
<td>1428 (97.9)</td>
<td>2368 (97.6)</td>
</tr>
<tr>
<td><strong>Living environment N (%)</strong></td>
<td>Suburban 2121 (42.5)</td>
<td>139 (41)</td>
<td>324 (42.1)</td>
<td>641 (43.9)</td>
<td>1017 (41.9)</td>
</tr>
<tr>
<td><strong>Relationship status N (%)</strong></td>
<td>Married/domestic partnership 2570 (51.5)</td>
<td>182 (53.7)</td>
<td>368 (47.9)</td>
<td>759 (52)</td>
<td>1261 (52)</td>
</tr>
<tr>
<td><strong>Sexually active N (%)</strong></td>
<td>Yes 3973 (79.6)</td>
<td>269 (79.4)</td>
<td>614 (79.8)</td>
<td>1176 (80.6)</td>
<td>1914 (78.9)</td>
</tr>
</tbody>
</table>

Table 1: Demographics and disease-related baseline information about the Phendo cohort and each short-term engagement group.
## Longitudinal engagement statistics.

<table>
<thead>
<tr>
<th>Distinct days tracked</th>
<th>Day-spans over which tracked</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Mean ± std</strong></td>
<td><strong>Mean ± std</strong></td>
</tr>
<tr>
<td>0%, 25%, 50%, 75%, 100%</td>
<td>0%, 25%, 50%, 75%, 100%</td>
</tr>
<tr>
<td>Regulars</td>
<td>136.39 ± 140.27</td>
</tr>
<tr>
<td>Usuals</td>
<td>38.1 ± 29.41</td>
</tr>
<tr>
<td>Occasionals</td>
<td>13.44 ± 13.52</td>
</tr>
<tr>
<td>Seldoms</td>
<td>5.02 ± 8.14</td>
</tr>
</tbody>
</table>

Table 2: Longitudinal engagement statistics over the Phendo app-usage timeline for each group within the Phendo cohort.

## Phenotype domains and questions

<table>
<thead>
<tr>
<th>Domain</th>
<th>Question</th>
<th>Answer type</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>How was my day</td>
<td>How was your day?</td>
<td>5 single-choice items</td>
<td>good, manageable, bad, unbearable</td>
</tr>
<tr>
<td></td>
<td>Daily journal</td>
<td>free text</td>
<td></td>
</tr>
<tr>
<td>How I self-manage</td>
<td>What did you do to self-manage?</td>
<td>14 multiple choice items</td>
<td>heat pack, massage, talk</td>
</tr>
<tr>
<td></td>
<td>Take any supplements?</td>
<td>user-specified in profile;</td>
<td>CBD oil (15 mg),</td>
</tr>
<tr>
<td></td>
<td>Did you do any of these exercises that hurt?</td>
<td>multiple choice</td>
<td>magnesium (500mg),</td>
</tr>
<tr>
<td></td>
<td>Did you do any of these exercises that help?</td>
<td>profile; multiple choice</td>
<td>running, sit-ups, lunges,</td>
</tr>
<tr>
<td></td>
<td>Did you eat any foods that worsen symptoms?</td>
<td>profile; multiple choice</td>
<td>kickboxing</td>
</tr>
<tr>
<td></td>
<td>Did you eat any foods that improve symptoms?</td>
<td>profile; multiple choice</td>
<td>sugar, gluten, white flour,</td>
</tr>
<tr>
<td></td>
<td>Take any hormones?</td>
<td>user-specified in profile;</td>
<td>beer</td>
</tr>
<tr>
<td></td>
<td>Which activities were hard to do?</td>
<td>multiple choice</td>
<td>fresh veggies, lean meat, nuts</td>
</tr>
<tr>
<td></td>
<td>Take any medication?</td>
<td>user-specified in profile;</td>
<td>progestin (implant), 1.5 mg</td>
</tr>
<tr>
<td></td>
<td>How was sex?</td>
<td>multiple choice</td>
<td>gabapentin (300 mg), bupropion (300 mg)</td>
</tr>
<tr>
<td></td>
<td>Do you have your period?</td>
<td>2 single choice items</td>
<td>painful during, painful after, avoided</td>
</tr>
<tr>
<td></td>
<td>Are you in pain now? (body location, intensity)</td>
<td>39 multiple choice items</td>
<td>ovaries, cramping</td>
</tr>
<tr>
<td>What happens to me</td>
<td>Any GI/Urine issues?</td>
<td>15 multiple choice items</td>
<td>endo belly, vomiting, constipation</td>
</tr>
<tr>
<td></td>
<td>Experiencing something else, other symptoms?</td>
<td>21 multiple choice items</td>
<td>fatigue, headache, swelling</td>
</tr>
<tr>
<td></td>
<td>How is your mood?</td>
<td>30 multiple choice items</td>
<td>calm, happy, angry, anxious</td>
</tr>
<tr>
<td></td>
<td>Are you bleeding?</td>
<td>3 multiple choice items</td>
<td>clots, spotting, breakthrough bleeding</td>
</tr>
</tbody>
</table>

Table 3: Phendo self-tracking domains, questions, types of answer, and example answers. Text in bold indicates shorthand for each question.
Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplementaryinformation.pdf