PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (**http://bmjopen.bmj.com/site/about/resources/checklist.pdf**) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

ARTICLE DETAILS

TITLE (PROVISIONAL)	An mHealth app using machine learning to increase physical activity in diabetes and depression: clinical trial protocol for the DIAMANTE Study
AUTHORS	Aguilera, A; Figueroa, Caroline; Hernandez-Ramos, Rosa; Sarkar, Urmimala; Cemballi, Anupama; Gomez-Pathak, Laura; Miramontes, Jose; Yom-Tov, Elad; Chakraborty, Bibhas; Yan, Xiaoxi; Xu, Jing; Modiri, Arghavan; Aggarwal, Jai; Jay Williams, Joseph; Lyles, Courtney

VERSION 1 - REVIEW

REVIEWER	Jonathan Rawstorn Institute for Physical Activity and Nutrition, Deakin University, Australia
REVIEW RETURNED	03-Jan-2020

GENERAL COMMENTS	1. There are several minor spelling and grammatical errors throughout. These will likely be addressed during copyediting so I will not itemise them here. 2. Could the authors please clarify the difference between active and passive PA monitoring specified for static and adaptive message groups? 3. The discussion states this study will address a lack of implementation in previous research but it is unclear whether this intervention is being implemented as part of routine clinical care, or whether the researchers are simply recruiting via primary care services. Can the authors expand their explanation of how this study is being implemented in clinical care? 4. Allocation concealment is not explicitly addressed, I suggest appending “thereby ensuring allocation concealment” or similar to the sentence “Patients will be automatically randomized into groups trough our secure server during onboarding of the app.” 5. The adaptive intervention design aims for a “just in time” message delivery schedule but outcomes are relatively stable measures that will not provide information about how participants respond to this message delivery schedule. Given the constant passive measurement of PA, have the authors considered exploring how PA behaviour changes acutely/adaptively after “just in time” message delivery? 6. Additional explanation of the adaptive learning algorithm would be useful. For example, is message delivery time matched to active times during recent days (i.e. reinforce/build on existing active behaviours) or inactive times during recent days (i.e. prompt change

	in sedentary behaviour)? How are the effects of feedback and motivational messages assessed to determine which are types of message are most effective? What role do are clinical/demographic characteristics play in adaptive message delivery when they remain static (or relatively static) throughout the intervention period? 7. There is no clear explanation of why researchers must be aware of the nature and frequency of messages participants receive (i.e. must be unblinded). A succinct explanation of why this design is necessary would help readers to understand why this potential source of bias cannot be minimised. 8. The first two sentences of the conclusion could easily be combined into a more succinct single sentence. 9. Figure 1 could be reorganised to integrate Figure 2, between the "RCT" and "6 month follow up" boxes in figure 1. 10. There is a significant discrepancy between the number of participants stated in the manuscript and the consent form. Is there a reason for this? 11. Re: reporting checklist: the protocol version number and date do not appear to be included?

REVIEWER	Hein de Vries Maastricht University, Netherlands
REVIEW RETURNED	03-Feb-2020

GENERAL COMMENTS	The paper describes an interesting idea focusing on analyzing the added effects of providing adaptive feedback in comparison to a standard text condition. Unfortunately the make-up of the standard text condition seems to be very different from the adaptive condition, thus not allowing clear conclusions as to whether potential differences between the conditions are due to the adaptive learning strategy, or due to the fact that the standard text condition onlyseems to provide feedback on PA and mood levels, but not on messages based on the COM-B model. Additionally, some descriptions may need a bit more elaboration/depth. Page 4 20-30 I would combine the first two paragraphs as they deal both about the seriousness of the problem and that something needs to be done. 32 Why particularly effective, are they more effective than others? Please compare effect sizes with for instance more regular f2f interventions. 36 70% of LI Americans (are these also including Latin Americans by the way?) have smartphones, so 30% cannot be reached. Is reach then higher or lower than through regular care? Contextualization here is needed. 41 Please explain safety net settings. Term may not be understood by non-US citizens 45 Personalization can be done in many different ways. More background information is needed here. There are ample examples of earlier work using computer tailoring techniques that also provide personalization. When introducing ML a comparison of both techniques may be relevant. Page 5 10 I miss something about the working mechanisms to validate the hypothesis. I expect that the authors expect that adaptive messaging will lead to more attention, appreciation and use and thus to better outcomes? Somewhat more background is needed here. There are

	theories that link personalization with usability and effectiveness which may be relevant to include here. 20 The rationale for formulating the secondary hypotheses are not provided in the intro. It is nice to have (so) many secondary hypotheses, but why do they authors think that these outcomes will occur? 49 What do the authors do to minimize risks of their not blinded design? Page 7 4 I would strongly recommend the authors to include process data, to be obtained via quantitative/qualitative methods, and to monitor usage items (e.g. how long they read messages, etc..). There is quite some literature out there describing the various techniques that can be used for obtaining more process evaluation data which is crucial to better understand mHealth. You also may find interaction effects. Page 8 and Page 9 The description of the content items of the intervention is rather vague. The COM-B model provides a generic framework, but its translation to practice can vary quite a lot from application to application. So which types of messages were generated, e.g. to change attitudes, self-efficacy, social support, action planning and goal setting? How many messages will they receive, which frequency. How is the wording and is that comparable to the other conditions, as you otherwise are comparing apples and oranges. Wording in the adaptive and non-adaptive version should be the same, as otherwise wording will be contaminated with personalization. For instance, if the wording in the adaptive version is done much more empathically than in the other text condition, it may be 'logical' that the adaptive version will do better. Also, in the adaptive version in principle the same amount of factors should be able to be addressed as in the text version. I take that the format of the text version will be having a fixed number of factors to be addressed, whereas the adaptive version may vary. But, are you starting of by using the same amount of factors and than for instance on the basis of likes (as in recommender systems) proceed as the result of the machine learning process? These elements are not well described in the document and really do need much more attention. Page 9 4 This text suggests that you will not be using the COM-B model for the static text messages but only feedback on PS and stress. Hence, you ARE comparing apples with oranges and this RCT is not going to be able to answer the question whether adaptive messaging is more effective than static messaging, as you do not provide statically texts with the same content. In the past there was a paper by Dijkstra et al (1998/1999?) that shows how a dismantling design can be set up to show the real added effects of tailoring. This study, however, seems to be contaminating adaptive messaging with content. 35 Please explain more what multi-armed bandit problems are, the standard reader may have no clue. 43 So the authors use a type of recommender technique? If so, please make this connection in the text when describing Thompson sampling. Page 11. 6. Do the authors only expect main effects or also interaction effects. For instance, it could be that adaptive tailoring is particularly suitable for lower motivated persons, as highly motivated persons may profit also from more static messages. 51 I think that estimating only 15% drop-out is overly optimistic.

	Drop-out rates of 50% or more are quite common in eHealth studies and patient studies. I fear that the study will be underpowered. Additionally, they posit an interaction effect due to language. Is the study sufficiently powered to detect those subgroup differences? Page 13 Some descriptions in the discussion are actually more background information that is better suited in the introduction. Discussions for protocol papers can be quite limited I think as there is not yet so much to be discussed. What could be discussed is the rationale for certain decisions made, which is done in the limitations section. 34. Due to the contamination, the study cannot draw conclusions as to why machine learning was really needed to increase relevance and efficacy of the messages, the structure of providing feedback is so very different in the static and adaptive arm. The static arm should also contain motivational messages similar to those in the adaptive arm, but than for instance provided at random, rather than adaptive. Only then one can conclude if adaptive learning via machine learning will add something. This limitation should be acknowledged.

VERSION 1 – AUTHOR RESPONSE

Reviewer(s)' Comments to Author:

Reviewer: 1
Reviewer Name: Jonathan Rawstorn
Institution and Country: Institute for Physical Activity and Nutrition, Deakin University, Australia
Please state any competing interests or state 'None declared': None declared

Dear Professor Rawstorn, thank you for your thorough comments which have helped us to improve our manuscript. We address your comments below.

Please leave your comments for the authors below
1. There are several minor spelling and grammatical errors throughout. These will likely be addressed during copyediting so I will not itemise them here.

Thank you. We have re-examined our manuscript for spelling errors and made improvements.

2. Could the authors please clarify the difference between active and passive PA monitoring specified for static and adaptive message groups?

There is no difference in the PA monitoring for the static and adaptive messaging groups. Both groups will be able to monitor their PA in the DIAMANTE app, which is passively collected by their phone pedometer, and receive daily messages with motivational content and feedback on their steps the previous day. Of note, we have now altered the static condition, and renamed it the 'uniform random condition'. We explain these changes in more detail in our response letter below. The difference between the groups is that the uniform random group might receive these messages with different frequencies, as the adaptive groups' messaging decision are driven by the learning algorithm and the uniform random group's messaging is predefined with equal probabilities to receive types of messages. The control group in contrast, does not receive any text messages with feedback on their steps. However, they will be able to go into the DIAMANTE app to monitor their daily steps, but we will not prompt them to do so.

We have now added the above explanation to the text in the manuscript, p. 8:

"Briefly, both the adaptive and uniform random group will receive the same types of messages: feedback (4 active categories plus no message) and motivation (3 active categories plus no message). However, the message categories, timing and frequency will be optimized by a reinforcement learning algorithm in the adaptive group, and will be delivered with equal probabilities in the uniform random group (following a uniform random distribution). The control group will not receive these messages (only a weekly mood check-in message). All groups will have the app downloaded on their phone and their steps will be passively tracked within the app. See Figure 1 for an overview."

3. The discussion states this study will address a lack of implementation in previous research but it is unclear whether this intervention is being implemented as part of routine clinical care, or whether the researchers are simply recruiting via primary care services. Can the authors expand their explanation of how this study is being implemented in clinical care?

We are using a blended approach. While the intervention is in addition to current care, there are ways we are attempting to make it more a part of patients clinical care. We now explain this is the manuscript, page 12:

"Here, we are using a blended design: while the intervention is in addition to current care, there are ways we are attempting to make it more a part of patients clinical care. For instance, patients are mainly approached through primary care health providers, who recommend eligible patients whom they think are directly interested in a physical activity intervention. In addition, we will make patients' data available to providers: a summary of the step increase for that patient at the conclusion of the study and updated PHQ-8 and GAD scores entered into the record. Our study therefore is the first step to addressing this gap because of its integration in primary care clinics that serve low-income patients. Future work should focus more specifically on implementation of the app as part of routine clinical care."

4. Allocation concealment is not explicitly addressed, I suggest appending "thereby ensuring allocation concealment" or similar to the sentence "Patients will be automatically randomized into groups through our secure server during onboarding of the app."

Thank you for this suggestion. We have made the appropriate change, page 4:

"Patients will be automatically randomized into groups through our secure server during onboarding of the app, hereby ensuring allocation concealment."

5. The adaptive intervention design aims for a "just in time" message delivery schedule but outcomes are relatively stable measures that will not provide information about how participants respond to this message delivery schedule. Given the constant passive measurement of PA, have the authors considered exploring how PA behaviour changes acutely/adaptively after "just in time" message delivery?

This is an interesting question that we will be able to assess in post hoc analyses that break down PA in different units of measurement (e.g. hourly). This type of breakdown can help us understand immediate impact of messages.

We elaborate on this in the discussion section, page 13:

"In exploratory post hoc analyses, we will also be able to examine the more immediate effect of physical activity messages, e.g. on hourly steps in addition to daily steps, which will help to improve

future physical activity interventions (deliver messages at the right times). For instance, it is possible that one could receive a message in the morning and make plans to walk in the afternoon or evening, or messages could have more of an immediate impact. This information is currently unknown."

6. Additional explanation of the adaptive learning algorithm would be useful. For example, is message delivery time matched to active times during recent days (i.e. reinforce/build on existing active behaviours) or inactive times during recent days (i.e. prompt change in sedentary behaviour)? How are the effects of feedback and motivational messages assessed to determine which are types of message are most effective? What role do are clinical/demographic characteristics play in adaptive message delivery when they remain static (or relatively static) throughout the intervention period?

Thank you. We now provide additional explanation in the methods section, page 9:

More specifically, each morning, the algorithm assesses which messages will likely increase the physical activity for every participant in the upcoming day, and at which time period this message should be delivered. The algorithm training data consists of all previously collected data of all participants (contextual variables), which includes the types of messages that were sent since start of the study, and within which time periods; daily physical activity (pedometer data), and select clinical/demographic data (such as age, language and depression scores) to improve prediction abilities.

The algorithm will rely on a Bayesian linear regression model with interactions to predict the change in activity from the previous day to the current day, considering each type of feedback/motivational text-message and period and time of the message is given, and the contextual variables.

7. There is no clear explanation of why researchers must be aware of the nature and frequency of messages participants receive (i.e. must be unblinded). A succinct explanation of why this design is necessary would help readers to understand why this potential source of bias cannot be minimised.

We now elaborate on this, page 5:

"Patients need to be informed of the nature and frequency of the messages they will be receiving and discuss this with investigators during the course of the study. Further, if messages are not being sent out appropriately, research assistants will contact the app developer to address errors within 24 hours. If physical activity data is not coming in, they may need to contact participants to ensure that the app is actively running on their phone, and assist participants with re-downloading the app if necessary. The necessity of these steps makes it unfeasible to blind the researchers."

8. The first two sentences of the conclusion could easily be combined into a more succinct single sentence.

Thank you, we have now made this change, page 12:

"In this randomized controlled trial, we aim to examine the effect of a smartphone app that uses reinforcement learning to predict the most effective messages for increasing physical activity in 276 low-income, ethnic and racial minority patients with diabetes and depression in urban public sector primary care clinics."

9. Figure 1 could be reorganised to integrate Figure 2, between the "RCT" and "6 month follow up" boxes in figure 1.

Thank you. We have combined figure 1 and figure 2 into one integrated figure.

10. There is a significant discrepancy between the number of participants stated in the manuscript and the consent form. Is there a reason for this?

We have now uploaded a new version of the consent form. This consent form states 240 participants, because this is the number when not accounting for drop-out. We are aiming to recruit 276 participants.

11. Re: reporting checklist: the protocol version number and date do not appear to be included?

Thank you. We have included the protocol version and number.

Reviewer: 2
Reviewer Name: Hein de Vries
Institution and Country: Maastricht University, Netherlands
Please state any competing interests or state 'None declared': None

Please leave your comments for the authors below
The paper describes an interesting idea focusing on analyzing the added effects of providing adaptive feedback in comparison to a standard text condition. Unfortunately the make-up of the standard text condition seems to be very different from the adaptive condition, thus not allowing clear conclusions as to whether potential differences between the conditions are due to the adaptive learning strategy, or due to the fact that the standard text condition only seems to provide feedback on PA and mood levels, but not on messages based on the COM-B model. Additionally, some descriptions may need a bit more elaboration/depth.

Dear Professor de Vries, thank you for your thoughtful and thorough review of our paper. We agree with your concerns and have altered the comparison condition (the standard texts or static) to be more comparable to the adaptive learning strategy (adaptive condition). Both now have the same make-up of messages but in the new comparison condition participants will receive the messages by a uniform random probability distribution. We have renamed the static condition 'uniform random condition'. Below we elaborate on our changes and respond to your specific comments.

Page 4
20-30 I would combine the first two paragraphs as they deal both about the seriousness of the problem and that something needs to be done.

We agree with this suggestion and have now combined the two paragraphs.

32 Why particularly effective, are they more effective than others? Please compare effect sizes with for instance more regular f2f interventions.

We agree that this is not accurately explained in the manuscript and have revised the text. Current mobile interventions are not necessarily more effective than face-to-face interventions. However, because of the ubiquity of mobile phones, effective mobile health application provides an opportunity to increase access to care for vulnerable populations and thereby decrease health care disparities. We have revised the text, page 3:

"Mobile applications have been found effective in helping patients engage in healthy behaviors including physical activity. For instance, a recent meta-analysis of nine RCT's concluded that smartphone apps that focus on physical activity have a moderate positive effect on increasing physical activity levels, and another meta-analysis including 18 studies moderate to large effect in daily step changes (12, 13). These effect sizes are similar to 'face-to-face' interventions(14).

However, because around 70% of lower income Americans currently own smartphones(15), and the ownership of smartphones is expected to increase in low income populations globally(16), mobile apps have great potential to reach individuals that normally do not have access to care. This can decrease existing decrease disparities in health."

36 70% of LI Americans (are these also including Latin Americans by the way?) have smartphones, so 30% cannot be reached. Is reach then higher or lower than through regular care? Contextualization here is needed.

The 70% indeed includes Latinos in the US. We expect the reach to be higher, because Latinos in the US particularly have lower access to care. We have added this information in the introduction, page 3:

"However, because around 70% of lower income Americans (including Latinos) currently own smartphones(15), and the ownership of smartphones is expected to increase in low income populations globally(16), mobile apps have great potential to reach individuals that normally do not have access to care. Further, mobile technology can help overcome existing barriers in access to care for vulnerable populations, including lower availability of psychological treatment in primary care settings, language and literacy barriers, stigma, cost, and inflexible employment schedules(17). Latinos in the US in particular show higher under-utilization rates of mental health treatment than non-Latino whites(18). Deploying effective mobile applications can therefore decrease existing decrease disparities in health."

41 Please explain safety net settings. Term may not be understood by non-US citizens

Thank you for pointing this out. We now explain this in the introduction, page 3:

"However, in the U.S., low-income minority patients frequently receive their care in safety net settings (services in the public sector for those unable to attain private health insurance), where novel mobile technologies are not often designed, developed or implemented(17)."

45 Personalization can be done in many different ways. More background information is needed here. There are ample examples of earlier work using computer tailoring techniques that also provide personalization. When introducing ML a comparison of both techniques may be relevant.
Page 5

We agree that this comparison is important. Here, personalization refers to a learning strategy to adapt the intervention based on predicted outcomes using historical participant data, which is more complex than earlier computer tailoring techniques. We have now altered the text accordingly, page 3-4:

"Further, most mobile applications that target behavioral changes are not personalized(18), which could contribute to lower effect sizes of these interventions for trials with longer study durations (e.g. over 3 months)(12). Smartphone interventions allow for data collection by passive sensing technologies, which offer an opportunity for tailoring and personalizing interventions to users behavior, preferences and needs. Personalization can be achieved by computer tailoring: tailoring interventions to observed behavior and characteristics of the participant. This can include feedback, goal setting, or user targeting (i.e. conveying that communication is designed specifically for the user)(19). However, more complex forms of personalization might be needed to increase and maintain engagement with PA interventions (20), a requirement for a digital intervention to be effective. One promising approach is to use adaptive learning, which allows the prediction of which content might be effective for users, learning from it's previous actions and participant data collected by mobile phones (21). Research on the efficacy of these treatments is still in its early stages."

10 I miss something about the working mechanisms to validate the hypothesis. I expect that the authors expect that adaptive messaging will lead to more attention, appreciation and use and thus to better outcomes? Somewhat more background is needed here. There are theories that link personalization with usability and effectiveness which may be relevant to include here.

Thank you for this comment. While engagement is indeed important, another working mechanism is matching motivation needs for each participant. We now explain this in the introduction, page 4:

“For instance, an adaptive learning algorithm might be more effective as it can match motivation needs for each participant. Some intervention studies have attempted cultural tailoring to large groups of people. However, most tailoring approaches occur at the group level whereas we are bringing the tailoring down to the individual level(24).”

20 The rationale for formulating the secondary hypotheses are not provided in the intro. It is nice to have (so) many secondary hypotheses, but why do they authors think that these outcomes will occur?

While PA is the primary outcome, it is conceptualized as being a mechanism for improving other health outcomes for diabetes and depression. Therefore, we would like to investigate the relationship of physical activity to the chain of variables leading to clinical outcomes, introduction, page 4:

“A growing body of evidence suggests that physical activity is such a risk factor: it is linked to both mental health and diabetic outcomes”

49 What do the authors do to minimize risks of their not blinded design?
Page 7

We believe that the automated and protocol based nature of our design and intervention has minimized risks of not blinding. First, participants are automatically assigned to a condition. Research Assitants will not know what condition the participant will be randomized in. Second, research assistants will not have contact with participants, unless they need to be called to restart their app. The protocol is highly standardized and technically focused for all arms of the study using a single procedure.

4 I would strongly recommend the authors to include process data, to be obtained via quantitative/qualitative methods, and to monitor usage items (e.g. how long they read messages, etc..). There is quite some literature out there describing the various techniques that can be used for obtaining more process evaluation data which is crucial to better understand mHealth. You also may find interaction effects.

Thank you for this suggestion. We are now providing more information on assessing engagement measures. Measures, page 7:

“Engagement measures
In addition to physical activity and the measures mentioned above, we will also examine engagement measures, such as (1) times that the app was opened, (2) time spent reading the messages, (3) usability data, assessed by the Systems Usability Scale(28) and (4) open ended qualitative questions enquiring participants opinions of the app.”

Page 8 and Page 9
The description of the content items of the intervention is rather vague. The COM-B model provides a generic framework, but its translation to practice can vary quite a lot from application to application. So which types of messages were generated, e.g. to change attitudes, self-efficacy, social support,

action planning and goal setting? How many messages will they receive, which frequency. How is the wording and is that comparable to the other conditions, as you otherwise are comparing apples and oranges. Wording in the adaptive and non-adaptive version should be the same, as otherwise wording will be contaminated with personalization. For instance, if the wording in the adaptive version is done much more empathically than in the other text condition, it may be 'logical' that the adaptive version will do better. Also, in the adaptive version in principle the same amount of factors should be able to be addressed as in the text version. I take that the format of the text version will be having a fixed number of factors to be addressed, whereas the adaptive version may vary. But, are you starting of by using the same amount of factors and than for instance on the basis of likes (as in recommender systems) proceed as the result of the machine learning process? These elements are not well described in the document and really do need much more attention.

Thank you for this comment. As mentioned earlier, according your comments we have scrutinized our static and adaptive arm and have now altered these arms to avoid contamination. Our adaptive arm and static arm (now called 'uniform random') will receive the same types of text messages, but the messages in the adaptive arm will be chosen by the reinforcement learning algorithm, and in the uniform random arm they will have equal probabilities (following a uniform random distribution).

Participants in both groups will thus receive up to two messages a day; a feedback (4 active categories plus no message) and a motivational message (3 active categories plus no message) within 4 possible time frames, selected from the same messaging banks. These messages are designed according to the COM-B model, a cognitive framework for behavior change. We have now altered this in the text, page 8:

"Briefly, both the adaptive and uniform random group will receive the same types of messages: feedback (4 active categories plus no message) and motivation (3 active categories plus no message). However, the message categories, timing and frequency will be optimized by a reinforcement learning algorithm in the adaptive group, and will be delivered with equal probabilities in the uniform random group (following a uniform random distribution). The control group will not receive these messages (only a weekly mood check-in message). All groups will have the app downloaded on their phone and their steps will be passively tracked within the app. See Figure 1 for an overview."

And page. 8:

"Uniform random message arm
We will send patients up to two messages per day within 4 randomly selected time intervals. These messages are based on the COM-B framework (examples shown in Table 1 A/B). In addition, they will receive one message, on the seventh day, which will ask patients to rate their mood on a scale from 1 to 9. Physical activity (step-count/day) will be passively monitored via the app on their smartphone.

Adaptive message arm
Patients in the adaptive messaging arm will receive the daily COM-B messages (equal to the uniform random arm, examples shown in Table 1 A/B), but the message categories, timing and frequency, will not be chosen randomly, but by using a reinforcement learning (RL) algorithm. This allows us to adequately assess whether differences in effects are driven by the use of the RL algorithm. Physical activity (step-count/day) will be actively monitored via the app on their smartphone."

Page 9
4 This text suggests that you will not be using the COM-B model for the static text messages but only feedback on PS and stress. Hence, you ARE comparing apples with oranges and this RCT is not going to be able to answer the question whether adaptive messaging is more effective than static

messaging, as you do not provide statically texts with the same content. In the past there was a paper by Dijkstra et al (1998/1999?) that shows how a dismantling design can be set up to show the real added effects of tailoring. This study, however, seems to be contaminating adaptive messaging with content.

Please see the responses to your previous comments.

35 Please explain more what multi-armed bandit problems are, the standard reader may have no clue.

We now provide more information about multi-armed bandits, page 9:

“A contextual MAB problem is a reinforcement learning setting, in which the algorithm choses between different treatment options which all have different reward functions. The reward functions depend on contextual variables.”

43 So the authors use a type of recommender technique? If so, please make this connection in the text when describing Thompson sampling.
Page 11.

Thompson sampling is an algorithm that can be used to analyze multi-arm bandit problems. As data accumulate on a participant, these algorithms increase the chance of providing the messages that are most effective in a particular situation, and decrease the chance of providing messages that are less effective. Thompson sampling can be used for recommender systems. In this case, we are not using a classical recommender system because patients are not providing their ratings or preferences for the messages. Rather, we use their observed physical activity as feedback for the learning algorithm.

We now provide more information about Thompson sampling in the manuscript, page 9:

More specifically, each morning, the algorithm assesses which messages will likely increase the physical activity for every participant in the upcoming day, and at which time period this message should be delivered. The algorithm training data consists of all previously collected data of all participants (contextual variables), which includes the types of messages that were sent since start of the study, and within which time periods; daily physical activity (pedometer data), and select clinical/demographic data (such as age, language and depression scores) to improve prediction abilities.

The algorithm will rely on a Bayesian linear regression model with interactions to predict the change in activity from the previous day to the current day, considering each type of feedback/motivational text-message and period and time the message is given, and the contextual variables.

6. Do the authors only expect main effects or also interaction effects. For instance, it could be that adaptive tailoring is particularly suitable for lower motivated persons, as highly motivated persons may profit also from more static messages.

Yes, these and other interaction effects are likely possible and can be assessed post-hoc, because interaction possibilities are almost limitless. Also, the algorithm will embed many interactions within its calculation which could account for some of these relationships. We decided to focus on main effects as the primary outcomes with the possibility of exploring the data in other ways after collection.

51 I think that estimating only 15% drop-out is overly optimistic. Drop-out rates of 50% or more are quite common in eHealth studies and patient studies. I fear that the study will be underpowered.

Additionally, they posit an interaction effect due to language. Is the study sufficiently powered to detect those subgroup differences?

Yes, we definitely understand that drop out in eHealth studies is high. We have a more optimistic 15% rate largely due to the fact that the primary outcome of interest will be passively assessed and thus requires limited proactive engagement by the user. We also have a comprehensive protocol to automatically text users who have closed the app or are not transmitting data. After an automated text, they will receive multiple calls to restart the app or receive technical assistance. We agree that drop out would be closer to 50% or higher if we simply had users download the app and not be contacted afterwards.

We now elborate on this in the text, page 11:

“We believe that low drop-out is feasible as (1) the primary outcome of interest will be passively assessed and thus requires limited proactive engagement by the user and (2) we have a comprehensive protocol to automatically text users who have closed the app or are not transmitting data.”

We did not power mainly for the subgroup analyses, since these are secondary, more exporatory evaulations. We expect at least 50% of the sample to be Spanish speakers given our previous testing recruitment. For secondary analyses examining step count differences within language groups, we expect to have sufficient power to detect larger changes (such as 2000 step count improvements), but did not power our recruitment strategy specifically for these post-hoc effects. Page 11:

“We will conduct an exploratory sub-group analyses for our primary and secondary hypotheses for English versus Spanish speaking patients. We expect at least 50% of the sample to be Spanish speakers given our previous testing recruitment.”

Page 13
Some descriptions in the discussion are actually more background information that is better suited in the introduction. Discussions for protocol papers can be quite limited I think as there is not yet so much to be discussed. What could be discussed is the rationale for certain decisions made, which is done in the limitations section.

Thank you. We have moved these descriptions to the introduction.

34. Due to the contamination, the study cannot draw conclusions as to why machine learning was really needed to increase relevance and efficacy of the messages, the structure of providing feedback is so very different in the static and adaptive arm. The static arm should also contain motivational messages similar to those in the adaptive arm, but than for instance provided at random, rather than adaptive. Only then one can conclude if adaptive learning via machine learning will add something. This limitation should be acknowledged.

We have changed the conditions (see answers to the previous comments).