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)	Vulnerability for New Episodes in Recurrent Major Depressive Disorder: Protocol for the Longitudinal DELTA-Neuroimaging Cohort Study
AUTHORS	Mocking, Roel; Figueroa, Caroline; Rive, Maria; Geugies, Hanneke; Servaas, Michelle; Assies, Johanna; Koeter, Maarten; Vaz, Frédéric; Wichers, Marieke; van Straalen, Jan; de Raedt, Rudi; Bockting, Claudi; Harmer, Catherine; Schene, Aart; Ruhé, Henricus

VERSION 1 - REVIEW

REVIEWER	Cédric Lemogne AP-HP, Université Paris Descartes, Inserm, France
REVIEW RETURNED	10-Apr-2015

GENERAL COMMENTS	Great protocol: The topic is of great interest The rationale is sound To recruit remitted unmedicated patients will be a strength of the study The multimodal MRI assessment is another strength The protocol is generally well described My main concern regards statistical power issues for testing hypotheses 3 and 4. This concern relates to the expected number of patients who will relapse within 2.5 years of follow-up. I am afraid that the study will lack statistical power to test predictors of recurence on one subsample (let say 30 patients with 15 experiencing relapse) and then confirm the predictive value of the model on the other subsample (Siegle et al. Arch Gen Psychiatry 2012) The authors write that: "Consequently, if we could lower recurrence rates in these recurring cases, we may greatly reduce the overall number of MDD-episodes and thereby MDD's burden." Although it might seem intuitive, the authors may want to cite evidence that the burden of MDD is alleviated between MDE (e.g. Judd et al., ArchGenPsychiatry 2000) The theoretical framework is nicely and concisely exposed. The authors may also want to integrate to their framework the increased attention to the self displayed by depressed individuals and how this increased self-focus relate to both affective, cognitive and brain features of both depressed and remitted patients (Nejad et al., Front Hum Neurosci 2013). This may also strengthen the rationale for using self-referent personality traits in the affective neuropsychological tests. The authors write that: "our own research indicates that HPA-axis hyperactivity is an endophenotypic trait"

	Is the term "endophenotypic" appropriate here? Unless there is evidence that unaffected first-degree relatives display the same pattern, "trait" might better stand alone. Furthermore, the authors may want to explain why the cortisol slope over the day (rather than absolute value at a given time) will not be assessed. The authors may want to provide more details about the recruitment of the patients and the way the diagnosis of recurring MDD will be ascertained. How past episodes of depression will be assessed? Will the SCID be used during an episode and the the patient followed up to remission to be included? Or will the SCID will be used only in a retrospective way? How the patients will be selected to be approached by phone? The choice of testing the reward system with non-social stimuli only (sweet or bitter drinks among thirsty participants) is questionable as differences in depressed individuals may be particularly pronounced for social stimuli (Davey et al.) The authors may want to further explain this choice. The authors may also want to provide the rationale that guided the order of the tasks during the MRI assessment (e.g. second resting state acquisition after and not before the emotion regulation task). The procedure by which patients experiencing a relapse will be matched with still-remitted patients is not described. What if this "control" subsequently experience relapse? Or if he or she is also a perfect match for another patient experiencing relapse later? I believe that this part is a bit confusing as it stands. Please consider a simplification of the protocol or doing a bit more to make the case for such design

REVIEWER	Jeremy G. Stewart, Ph.D. McLean Hospital and Harvard Medical School United States
REVIEW RETURNED	11-May-2015

GENERAL COMMENTS	The proposed study will employ an extensive, multi-method longitudinal assessment of adults diagnosed with recurrent (at least 2 previous episodes) MDD. The general aim of the project is to develop a better understanding of MDD recurrence at multiple units of analysis (i.e., symptoms, affective neuropsychology, brain circuitry, endocrinology/metabolism). The proposed project is innovative, extremely rigorous and data-rich, and has a number of strengths, which the authors enumerate. Please see comments below that correspond to items of the review checklist to which I responded 'No'. 3) Is the study design appropriate to answer the research question? a) I think the authors should expand on their decision to focus this study on a subset of individuals who have 2 or more previous episodes of depression. By definition, this subset of individuals is already highly recurrent. As Monroe and Harkness (2011; cited by the authors) discuss, a substantial portion (40-50%) of individuals suffering from MDD ONLY have one episode. The authors propose

	to examine predictors of recurrence in a sample of depressed adults with an extremely high likelihood of experiencing recurrence (approximately 90% of individuals with at least 2 previous episodes; see Harkness & Monroe, 2011, Figure 1). Because of the sample, the project is not in a position to identify factors that differentiate individuals who have a non-recurrent course versus those who have a recurrent course. Which, some might argue, would say more about the potential mechanisms that explain or drive recurrence in MDD. Instead, the project will provide insight on the factors associated with a course of MDD that is characterized by many, many episodes, as opposed to only a few episodes. I think that this warrants further consideration and comment. At the very least, this should be discussed as a limitation of the proposed project, and some of the wording surrounding what the results might mean for understanding recurrence could be altered accordingly. As a suggestion, the authors could consider recruiting a sample of adults who report Single Lifetime Episodes of Depression (SLEDs; Monroe & Harkness, 2011) instead of the control subjects at baseline. This would allow the authors to make more specific inferences regarding the factors that are specifically tied to recurrence. b) Participants are excluded from participation if they are using psychoactive medications. However, the authors do not describe their approach to assessing and handling other types of active treatments (e.g., empirically supported psychotherapy; unstructured counseling; support groups). Given active psychotherapy would surely have an impact on recurrence, the authors should specify how this will be handled in the current protocol. c) Unless I am mistaken, it appears the authors propose having the participants eat and/or drink after the blood draw on the baseline visit. On that same day, the authors collect saliva samples that they will ultimately assay for cortisol. My understanding is that participants should not eat or drink (aside from water) several hours before providing saliva samples. d) The authors appear to intend to examine HPA axis response (measured through salivary cortisol) to their sad mood induction in the first session. Could the authors provide more information on their justification for doing so, and what they expect to find. Emotion inductions (such as the sad mood induction the authors describe) do not produce a robust HPA axis response (Dickerson & Kemeny, 2004). 5) Are research ethics (e.g. participant consent, ethics approval) addressed appropriately? a) The authors do not speak about any specific procedures that may be in place for patients who express acute risk for suicidal behaviors during the follow-up period. Items in the questionnaire packets would measure suicide directly. Are there any procedures for contacting the patients and/or their outpatient treaters (if applicable) in these cases? If so, what is the threshold at which interventions are necessary? 7) If statistics are used are they appropriate and described fully? a) Could the authors provide more specific detail about the following

	: i) the analyses pertaining to the ESM data and ii) the analyses for the third and fourth hypotheses (particularly commenting on handling time invariant and time variant covariates in the cox regression models). b) This protocol should be reviewed by a statistical specialist, as I've indicated.

REVIEWER	William Iacono University of Minnesota, USA
REVIEW RETURNED	19-May-2015

GENERAL COMMENTS	This study of factors that predict recurrent depression (MDD) offers a comprehensive assessment protocol using interesting assessments to tackle this important public health matter. The proposal is generally well written. What is missing for me is enough elaboration of the rationale for the study design for me to see why the project needs to be structured the way it is or how the current design is somehow optimal. At best I could see this as an interesting pilot study. A major concern is that it is almost certainly underpowered at 60 participants, and there is no allowance for attrition. The explanation that power analyses of MRI studies are not needed, besides being arguable, omits consideration of the fact that most of the measures are not MRI based. There are many measures, making false positive outcomes a distinct possibility. Gender is a major risk factor for MDD, and its role in recurrent MDD is thus of considerable interest, but gender is not mentioned in the proposal. The sample is not likely to be in any way representative of those with recurrent MDD because the participants cannot be in treatment and are to be free of comorbidity. Unfortunately, the comorbidity is likely to be important to relapse. The value of the control group is not clear. Comparing recurrent MDD cases to never depressed super healthy controls does not make possible separation of effects reflecting risk for MDD, risk for MDD recurrence, consequences of MDD, or consequences of MDD recurrence. These effects are not mutually exclusive, but these investigators needed to describe how they isolate the key effects of interest. Much of the work in this project is devoted to follow-up assessment that involves e.g., re-scanning the patients, but it is not evident how scanning the patients when they have a relapse using a complex and expensive design, is worth the effort given the key hypotheses, especially since it will be difficult to control medication status at the time of relapse. It is important to know who relapses, but the risk factors will be evident by contrasting relapsers to non-relapsers on their baseline measures. I also did not understand how recruiting a mixed MDD sample of people through newspaper ads and from the Delta study was to be accomplished such that it would be easy to interpret the results since those in the Delta study are being treated with preventive cognitive therapy to lessen the likelihood of recurrence. This seems to introduce a confound that further complicates this small N study.

VERSION 1 – AUTHOR RESPONSE

Great protocol:
The topic is of great interest
The rationale is sound

To recruit remitted unmedicated patients will be a strength of the study The multimodal MRI assessment is another strength The protocol is generally well described

1. My main concern regards statistical power issues for testing hypotheses 3 and 4. This concern relates to the expected number of patients who will relapse within 2.5 years of follow-up. I am afraid that the study will lack statistical power to test predictors of recurrence on one subsample (let say 30 patients with 15 experiencing relapse) and then confirm the predictive value of the model on the other subsample (Siegle et al. Arch Gen Psychiatry 2012)

We thank Reviewer #1 for expressing and explaining his/her concerns regarding statistical power. This refers to hypotheses 3 (In remitted unmedicated recurrent MDD-subjects, the above [neuropsychological and brain function] alterations in will predict prospective 2.5yrs follow-up symptom course) and 4 (above [neuropsychological and brain function] alterations will become more pronounced during repeated measures in recurrent MDD-subjects experiencing a recurrence during follow-up, in comparison to repeated measures in matched remitted recurrent MDD-subjects).

We agree with Reviewer #1 that power calculations for imaging studies remain hard, and many studies are underpowered due to the costs associated with imaging research. Interestingly, Reviewer #1 cites a study that is comparable to our study regarding power analysis. In that study, 49 unmedicated depressed adults and 35 healthy control participants were included from two separate studies, which provided adequate power for testing the stated hypothesis (Siegle et al., Arch Gen Psych, 2012). Of note, this is a smaller sample than the 60 patients and 40 controls in our study protocol. Moreover, Siegle et al. combined two cohorts, which may have introduced noise in the data which potentially reduced power. In addition, the distribution of outcomes in Siegle et al.'s study was about 2/3 vs. 1/3, while our expected distribution is 50-50 (1/2 vs. 1/2), which may have resulted in lower power than in our study. Therefore, in conclusion, based on these parameters we think the power of our study is at least equal, but likely larger than the study of Siegle et al., that was cited by Reviewer #1. In addition, more advanced hypothesis testing procedures (e.g. leave one out cross-validation, better distribution between train and test samples) would allow for more power that the procedure suggested by Reviewer #1. Finally, cooperating with other studies to obtain external validation samples has been proven a legitimate way to validate results.

2. The authors write that: "Consequently, if we could lower recurrence rates in these recurring cases, we may greatly reduce the overall number of MDD-episodes and thereby MDD's burden." Although it might seem intuitive, the authors may want to cite evidence that the burden of MDD is alleviated between MDE (e.g. Judd et al., ArchGenPsychiatry 2000)

We thank Reviewer #1 for this suggestion to support intuition with evidence.

3. The theoretical framework is nicely and concisely exposed. The authors may also want to integrate to their framework the increased attention to the self displayed by depressed individuals and how this increased self-focus relate to both affective, cognitive and brain features of both depressed and remitted patients (Nejad et al., Front Hum Neurosci 2013). This may also strengthen the rationale for using self-referent personality traits in the affective neuropsychological tests.

We agree with Reviewer #1 that negative biases in depression also pertain to the self. In our protocol we focus on the self in the various affective neuropsychological tasks that use self referent negative and positive emotional words, and in measures of self-referential processing as rumination and cognitive reactivity. Furthermore, in our protocol we focus on activity and connectivity of resting-state networks as the Default Mode Network, which has a significant role in self-associative and introspective processes and overlaps with import cortical midline regions involved in self-referential processing mentioned by Nejad and colleagues. Our measures of self-referential processing can be

related to resting-state connectivity measures of the DMN and other relevant neural networks. We would be happy to incorporate the suggested reference to further support our theoretical framework.

4. The authors write that: "our own research indicates that HPA-axis hyperactivity is an endophenotypic trait". Is the term "endophenotypic" appropriate here? Unless there is evidence that unaffected first-degree relatives display the same pattern, "trait" might better stand alone.

We agree with Reviewer #1 that unaffected first-degree relatives showing the same pattern would support the concept of an endophenotype as proposed by Gottesman & Gould. Indeed, in the referenced papers, this evidence is provided in first degree relatives as suggested by Reviewer #1. Of note, we write that "Also our own research indicates that HPA-axis hyperactivity is an endophenotypic trait", by which we mean that our research in conjunction with other findings supports the endophenotype concept for HPA-axis hyperactivity in depression.

5. Furthermore, the authors may want to explain why the cortisol slope over the day (rather than absolute value at a given time) will not be assessed.

We agree with Reviewer #1 that the cortisol slope over the day provides important information regarding HPA-axis activity. Therefore, as we intended to describe in the supplemental material, we propose to measure 5 cortisol measures over the day, including a measurement at 22:00h. This way, we can model the diurnal cortisol slope over the day, as suggested by Reviewer #1. This description could be further clarified.

6. The authors may want to provide more details about the recruitment of the patients and the way the diagnosis of recurring MDD will be ascertained. How past episodes of depression will be assessed? Will the SCID be used during an episode and [is] the patient followed up to remission to be included? Or will the SCID will be used only in a retrospective way? How the patients will be selected to be approached by phone?

We agree that the description of our recruitment and diagnostic procedures could be clarified. In detail, we will use the SCID both cross-sectional, retrospective and prospective. At the moment of potential inclusion, we use the SCID to verify that the subject is in remission at the time of inclusion. Moreover, we retrospectively verify past episodes. Subsequently, we follow-up the patients with regular SCID assessments to cross-sectionally test for depressive episodes during the assessment and retrospectively since the last assessment. Of note, the SCID is widely validated (also telephonically) and considered the golden standard of depressive episode diagnosis, also retrospectively. While the design that Reviewer #1 proposes (following patient up to remission) would be interesting, this would lead to inclusion of patient that are only in remission very briefly. This would strongly reduce the generalizability of findings, which could also be a caveat for our research questions. Regarding the selection of patients to be approached by phone, we follow a combined and open recruitment strategy in order to minimize selection biases. In detail, we recruit both groups through identical advertisements in freely available online and house-to-house papers, posters in public spaces, from previous clinical studies in our and affiliated research centres, and from patients previously treated in the AMC or affiliated general practitioners and psychologists.

7. The choice of testing the reward system with non-social stimuli only (sweet or bitter drinks among thirsty participants) is questionable as differences in depressed individuals may be particularly pronounced for social stimuli (Davey et al.) The authors may want to further explain this choice.

We agree with reviewer #1 that the test battery for the reward system could have been extended in order to have more data of e.g. social rewarding stimuli. This might be covered partly by the ESM-data in our sample, but is definitely not assessed like in the work of Davey et al (2010; 2011). However as the primary focus of this research was based on cognitive reactivity and emotional biases we could only investigate reward function in a limited way. In this we choose to investigate the persistence of difficulties in temporal difference reward related learning with primary rewards, as this could be a more general and basic persistent dysfunction in recurrent MDD.

8. The authors may also want to provide the rationale that guided the order of the tasks during the MRI assessment (e.g. second resting state acquisition after and not before the emotion regulation task).

Due to time constraints we had to perform the emotion regulation task (ERT) in the second block of scanning, after the break. It could be questioned whether the brain activation by this ERT might be recognized in the 2nd resting state scan after a mood-induction. If anything, we aimed to have the subjects maximally experience a sad mood after the mood-induction. As the ERT also provided negative pictures, (alternated with positive) it can be expected that the ERT might also have primed people to be more susceptible for the mood induction procedure. As this was a systematic order of scanning in all subjects, we think that if any effect occurred, this would have only primed all subjects systematically to be more vulnerable for the mood induction.

9. The procedure by which patients experiencing a relapse will be matched with still-remitted patients is not described. What if this "control" subsequently experience relapse? Or if he or she is also a perfect match for another patient experiencing relapse later? I believe that this part is a bit confusing as it stands. Please consider a simplification of the protocol or doing a bit more to make the case for such design

We agree with Reviewer #1 that the explanation of this follow-up matching procedure could be clarified. In detail:

For the prediction of recurrences every participant will be followed-up for 2.5 years and all recurrences over these 2.5 years and their duration will be identified/quantified. When patients experience a relapse and agree to participate in the study again, they will indeed be matched with participants who are in remission at that time, and also meet matching criteria (follow -up time, age, sex, educational level and working class). These matched participants have to be currently euthymic but can have had a prior relapse, thus after the baseline measurement, or a relapse during follow-up, after second participation. The reason for this approach is that we are interested in comparing the effect of depression (state) vs. depressive vulnerability (trait), instead of simply comparing more vulnerable patients to stable patients. This will give us insight into the pathophysiology of relapse vs. remission; which factors stay the same, and which factors show change when patients relapse. Potential in-between recurrences will, however, be examined as a potential confounder in the final analyses. Nevertheless a participant will be never included more than once in the follow-up repetitive measurements (scanning/neuropsychology), in order to exclude the possibility of learning effects and habituation in testing/scanning and prevent complex covariance structures.

To further clarify the procedure, we conducted matching based on group-level characteristics of relapse patients vs. control patients (mean follow-up time, age, years, sex, educational level and working class). This avoids issues such as posed by reviewer#1, as control patients don't have to be perfect matches for one participant, but have to match on a group level (distribution: average and spread). Thus, if a relapsed patient participates in the study for a second time, we examine how differences in group characteristics between the relapse group and the control group would statistically change when including particular controls. In our selection of control participants, we aim

to invite controls with characteristics that keep groups matched, thus aiming to maintain an optimal balance between the relapse group and the control group. In this way, we also aim to include relatively more controls, (relapse:control=1:1.5), with the goal of increasing power.

Reviewer #2

Reviewer Name: Jeremy G. Stewart, Ph.D.
Institution and Country: McLean Hospital and Harvard Medical School

The proposed study will employ an extensive, multi-method longitudinal assessment of adults diagnosed with recurrent (at least 2 previous episodes) MDD. The general aim of the project is to develop a better understanding of MDD recurrence at multiple units of analysis (i.e., symptoms, affective neuropsychology, brain circuitry, endocrinology/metabolism).

The proposed project is innovative, extremely rigorous and data-rich, and has a number of strengths, which the authors enumerate. Please see comments below that correspond to items of the review checklist to which I responded 'No'.

3) Is the study design appropriate to answer the research question?

a) I think the authors should expand on their decision to focus this study on a subset of individuals who have 2 or more previous episodes of depression. By definition, this subset of individuals is already highly recurrent. As Monroe and Harkness (2011; cited by the authors) discuss, a substantial portion (40-50%) of individuals suffering from MDD ONLY have one episode. The authors propose to examine predictors of recurrence in a sample of depressed adults with an extremely high likelihood of experiencing recurrence (approximately 90% of individuals with at least 2 previous episodes; see Harkness & Monroe, 2011, Figure 1). Because of the sample, the project is not in a position to identify factors that differentiate individuals who have a non-recurrent course versus those who have a recurrent course. Which, some might argue, would say more about the potential mechanisms that explain or drive recurrence in MDD. Instead, the project will provide insight on the factors associated with a course of MDD that is characterized by many, many episodes, as opposed to only a few episodes. I think that this warrants further consideration and comment. At the very least, this should be discussed as a limitation of the proposed project, and some of the wording surrounding what the results might mean for understanding recurrence could be altered accordingly.

As a suggestion, the authors could consider recruiting a sample of adults who report Single Lifetime Episodes of Depression (SLEDs; Monroe & Harkness, 2011) instead of the control subjects at baseline. This would allow the authors to make more specific inferences regarding the factors that are specifically tied to recurrence.

We thank Reviewer #2 for his/her detailed analyses of the course of depression and its consequences for our research design. While we agree that the idea of Monroe and Harkness that a substantial portion of individuals that experienced an MDD episode will only have one episode is very hopeful, it is not a generally accepted truth. In their paper, Monroe and Harkness make a compelling case to study subjects with single lifetime episodes of depression. However, in the end, the only way to make sure that a subject had only one lifetime episode of depression would mean a lifetime follow-up from birth to death, which would be costly, practically impossible and questionable from an ethical perspective. Interestingly, Monroe and Harkness do not cite the most extensive epidemiological study in depression performed to date, the International Consortium of Psychiatric Epidemiology (ICPE) Surveys (Andrade et al., 2003). In a total sample size of more than 37,000 subject around the world, recurrent episodes were reported by nearly 75% of respondents. Of note, this number could only increase because the subjects remain at risk in the time between assessment and death. This large dataset provides a less optimistic view on the course of depression. Of course the precise percentage

could be debated, but the view of Monroe and Harkness is generally considered as being on the more positive end of the spectrum. Nevertheless, it would be very interesting to include subjects that have only one lifetime episode, however with recurrence percentages slightly higher than the ones suggested by Monroe and Harkness, this would require very extensive research efforts including large samples and very long-term follow-ups.
The main focus of our project is the clinically relevant question to better predict what the individual risk of recurrence of a patient with recurrent MDD will be in the coming years, and how soon a recurrence will occur. From that perspective, the distinction with single lifetime episodes vs. recurrences will be of additional importance only. By inclusion of patients with less (≥2) and many recurrences we will have a spread of vulnerability, exactly as the reviewer suggests ("the factors associated with a course of MDD that is characterized by many, many episodes, as opposed to only a few episodes."). If we succeed, we would be able to select subjects at high risk of recurrence that need more aggressive preventive measures now to prevent imminent recurrence.

In line with this aim, we chose time to recurrence as our primary outcome measure, as opposed to a binary definition of recurrence yes vs. no. In addition, the group that could benefit most from preventive measures would likely be a group with a high priori chance, that is why we chose to recruit a sample with at 2 previous episodes or more. Of note, following figure 1 in Monroe and Harkness, this group has a 70% chance of recurrence during lifetime, of which we have previously shown that about 50% recurs in the first two years. In our group, there is also a spread in vulnerability, as there is a large spread in the number of previous episodes (e.g. from 2 up to 60), and the risk for recurrence greatly increases with number of episodes. Second, a substantial percentage of our remitted patients have had no or only 1 recurrence in the last 10 years. Some of these patients might be less vulnerable to recurrence than patients that have recently suffered from recurrences, as the risk of recurrence progressively decreased as the duration of recovery increases (Solomon et al, 2000). From these data, we think that our group would be the most interesting and feasible group to study when looking for factors that can predict imminent recurrence, in order to (I) select subjects that may benefit from preventive treatment, and (II) identify pathophysiological mechanisms that can be targeted in these subjects to prevent recurrence risk. We agree with Reviewer #2 that this could be better explained in the manuscript, which we can amend.

b) Participants are excluded from participation if they are using psychoactive medications. However, the authors do not describe their approach to assessing and handling other types of active treatments (e.g., empirically supported psychotherapy; unstructured counseling; support groups). Given active psychotherapy would surely have an impact on recurrence, the authors should specify how this will be handled in the current protocol.

We agree with Reviewer #2 that other therapies may have an influence on recurrence. However, it would be not feasible to exclude all other therapies as well, as we expect that it will be very hard to include medication free subjects already, including subjects without any form of therapy will be almost impossible. However, we will assess all forms of therapy used, report these and treat them as covariates in our analyses. Our current experience is that about 80% of the included patients do not receive any form of psychotherapy at baseline

c) Unless I am mistaken, it appears the authors propose having the participants eat and/or drink after the blood draw on the baseline visit. On that same day, the authors collect saliva samples that they will ultimately assay for cortisol. My understanding is that participants should not eat or drink (aside from water) several hours before providing saliva samples.

We thank Reviewer #2 for expressing and explaining his concerns regarding the saliva sampling procedure. The standard for obtaining saliva samples for cortisol analysis is that participants do not

eat/drink 15min before sampling, including water, which is in line with our protocol. Of note, the saliva is collected at a different day than the blood-draw.

d) The authors appear to intend to examine HPA axis response (measured through salivary cortisol) to their sad mood induction in the first session. Could the authors provide more information on their justification for doing so, and what they expect to find. Emotion inductions (such as the sad mood induction the authors describe) do not produce a robust HPA axis response (Dickerson & Kemeny, 2004).

We thank Reviewer #2 for the suggestion of the very interesting paper by Dickerson & Kemeny. Although the study indicates that, on average, the emotion induction stressor did not elicit a significant cortisol response, this should be interpreted with caution because of the relatively small numbers of studies that fell in this category as acknowledged by the authors. Moreover, it could be that some studies observed a positive, and others a negative effect, that levels out as no effect overall. In addition, Dickerson & Kemeny excluded studies in which recruitment was based on a physical or psychological diagnosis or a stressful experience (e.g., diabetes, depression, bereavement). This makes it hard to extrapolate their findings to our sample of recurrently depressed patients. Of note, several more recent papers did observe interesting effect of mood on salivary cortisol in recurrent depression (Chopra et al., 2008; Huffziger et al., 2013), which makes this assessment of great interest to our study.

5) Are research ethics (e.g. participant consent, ethics approval) addressed appropriately?

a) The authors do not speak about any specific procedures that may be in place for patients who express acute risk for suicidal behaviors during the follow-up period. Items in the questionnaire packets would measure suicide directly. Are there any procedures for contacting the patients and/or their outpatient treaters (if applicable) in these cases? If so, what is the threshold at which interventions are necessary?

We will contact participants regularly during follow-up. We will then address suicidality if this occurs during the telephone interviews and/or when a patient returns for a second assessment. As we are not the treating physician for these patients we will primarily advise patients to seek help for suicidal ideation if this responsibility is deemed acceptable. However, for emergency situations and/or imminent threat of suicidality we have a psychiatrist available for consultation. If necessary, we will refer a patient to the most appropriate emergency service (e.g. at our hospital or the most nearby hospital). We will then also inform the treating physician/GP afterwards. This procedure was approved by the research ethics committee.

7) If statistics are used are they appropriate and described fully?

a) Could the authors provide more specific detail about the following : i) the analyses pertaining to the ESM data and ii) the analyses for the third and fourth hypotheses (particularly commenting on handling time invariant and time variant covariates in the cox regression models).

i) Regarding the ESM a standard approach for data cleaning will be used. We will first check for missing data. Second, we will check whether total response time exceeds 15 minutes and whether time between the beep and first response exceeds 15 minutes, which observations will be removed. Third, we will excluded days of measurement when the number of observations was less than five. Fourth, we will exclude subjects when the number of observations is less than 30. These precautions are taken to have enough and valid measurements, necessary for valid statistical approaches. We will thereafter inspect the variables to see whether they contain variation based on the interquartile range.

Because ESM observations are irregularly spaced (due to the random presentation of measurements and missing data) and a positive/negative autocorrelation may exist between the expected absolute successive difference (EASD) and time intervals, we will calculate the mean adjusted absolute successive difference (MAASD) per ESM variable, taking into account an adjustment parameter λ, to capture affective instability (Jahng et al. 2008). To avoid night time intervals, successive differences will be calculated within days.

Because ESM-data will likely be skewed to the left, we will apply nonparametric independent samples Mann-Whitney U test to determine significance of differences between the remitted recurrent MDD and healthy control groups.

ii) We agree with Reviewer #2 that the description of our cox-regression models could be elaborated more. In detail, in first instance we are planning to only use the baseline predictors which are per definition time invariant. However, in a later stage, it would indeed be interesting to incorporate the variables that we measure over time, e.g. the HDRS or rumination questionnaires, to see how changes in these parameters over time are associated with future recurrence (mediation) and/or time until recurrence.

b) This protocol should be reviewed by a statistical specialist, as I've indicated.

This protocol has been written, and the study will be carried, out under close supervision of a statistical specialist, Dr. M.W.J. Koeter. We will describe this more clearly in our protocol.

Reviewer #3

Reviewer Name: William Iacono
Institution and Country: University of Minnesota

This study of factors that predict recurrent depression (MDD) offers a comprehensive assessment protocol using interesting assessments to tackle this important public health matter. The proposal is generally well written. What is missing for me is enough elaboration of the rationale for the study design for me to see why the project needs to be structured the way it is or how the current design is somehow optimal. At best I could see this as an interesting pilot study.

1. A major concern is that it is almost certainly underpowered at 60 participants, and there is no allowance for attrition. The explanation that power analyses of MRI studies are not needed, besides being arguable, omits consideration of the fact that most of the measures are not MRI based. There are many measures, making false positive outcomes a distinct possibility.

As shown in the appendix, we performed power-analyses for several outcomes, for continuous and categorical data. These power analyses showed adequate power to determine small to medium effect-sizes. Furthermore , our sample size is much larger than many previous studies. E.g. the study by Leyman et al. 2007 determined attentional biases with an exogenous cueing task in 20 MDD-patients versus 20 healthy controls.

We agree with the reviewer that we perform a large set of measurements, which carries the risk of false positives. However, as we will perform analyses accordingly to analysis-plans which will be a priori specified , we will do so for independent a priori hypotheses. Nevertheless this is the risk of a large endeavor like this study or for example the Netherlands Study for Depression and Anxiety where a group of patients is phenotyped well, which is also a strength of this study.

To further reduce the risk of chance findings we will use multivariate machine-learning approaches when larger numbers of (comparable) outcomes are considered, for example when we will analyse the facial expression recognition task. This task was recently tested to predict outcome of

antidepressant treatment in 60 patients treated with escitalopram. With a support vector machine learning approach, measurements at 2 weeks were predictive of antidepressant response at 8 weeks with accuracies around 75% (Browning, pers. Communication 2015).

Nevertheless, although our sample size will exceed the level of a pilot-study, especially for the prediction measures that we will identify, we will need new samples to replicate our findings.

Fortunately the attrition rate in the present study (and in the previous DELTA sample) is very low until now, which represents a high awareness of the necessity of this type of research amongst participants. In addition, all participants can be included in the cox-regression analyses, since these can adequately deal with attrition (outcome measure incorporates time to event or end of observation). Together with our power calculations, we are therefore less concerned of attrition than reviewer #3 might be.

2. Gender is a major risk factor for MDD, and its role in recurrent MDD is thus of considerable interest, but gender is not mentioned in the proposal.

Gender has often been investigated in MDD but was consistently found not to be a risk factor for recurrence, as was stated in the manuscript: demographics generally do not predict recurrence[18] (p5). We will nevertheless test associations with gender in our analyses and discard the variable from the models when gender is not a significant confounder.

3. The sample is not likely to be in any way representative of those with recurrent MDD because the participants cannot be in treatment and are to be free of comorbidity. Unfortunately, the comorbidity is likely to be important to relapse.

Comorbidity with e.g. anxiety disorders is no exclusion criterion, although primary anxiety disorders will be excluded (p12). The latter will keep homogeneity of the disorder (recurrent MDD) in the studied sample, which will further increase our power. We will determine comorbidity with the SCID, and take this outcome into account as a potential predictor.

4. The value of the control group is not clear. Comparing recurrent MDD cases to never depressed super healthy controls does not make possible separation of effects reflecting risk for MDD, risk for MDD recurrence, consequences of MDD, or consequences of MDD recurrence. These effects are not mutually exclusive, but these investigators needed to describe how they isolate the key effects of interest.

In this study, we compare remitted MDD-patients with recurrent episodes versus healthy controls. This contrast will address vulnerability factors for recurrence of MDD. We agree with the reviewer that associations with recurrence can be causal factors, consequences or confounders. The reviewer is right that we cannot disentangle these, but by comparing these groups, we maximize the contrast to have optimal power to detect these vulnerability factors. When we can demonstrate that these factors are associated with recurrence-rate, this could be indicative of causality but also of consequences. Potential confounding will be addressed by multivariate models.

See also the response to reviewer #2 point 3a.

5. Much of the work in this project is devoted to follow-up assessment that involves e.g., re-scanning the patients, but it is not evident how scanning the patients when they have a relapse using a complex and expensive design, is worth the effort given the key hypotheses, especially since it will be difficult

to control medication status at the time of relapse. It is important to know who relapses, but the risk factors will be evident by contrasting relapsers to non-relapsers on their baseline measures.

We agree with the reviewer that by repetitive scanning, we will not address research questions regarding risk factors for prediction of recurrence. With the repeated scans, we will be able to compare psychological measures and brain function after and during a depressive episode (Hypothesis 4; p11). This will provide important and innovative information regarding the pathophysiological changes during a relapse. This type of data is hardly existing to date. Our cohort will enable us to nicely match repetitive scanning and time-effects because we will also scan patients who do not have a recurrence (matched for follow-up time, age, sex, educational level and working class). This unique data merits the investment required for these additional measurements.

6. I also did not understand how recruiting a mixed MDD sample of people through newspaper ads and from the Delta study was to be accomplished such that it would be easy to interpret the results since those in the Delta study are being treated with preventive cognitive therapy to lessen the likelihood of recurrence. This seems to introduce a confound that further complicates this small N study.

The reviewer rightly points at an important issue that merits further explanation. Indeed an approximately expected 30% of patients participated in the original DELTA, >10 years before, of whom approximately 50% received randomized Preventive Cognitive Therapy (PCT). Although it is supportive for the strength of psychotherapy and PCT in specific, we think that the confounding effect of this previous treatment will not be huge. Furthermore, as previous psychotherapy (i.e. cognitive behavioral therapy, interpersonal Psychotherapy, PCT) was not an exclusion criterion in the total sample and the PCT intervention was more widely implemented in the Netherlands after 2005, this confounder will also exist in other participants (non-DELTA). We collect data on previous treatments in all subjects and will be able to estimate the magnitude of this possible confounding. Of note, the original DELTA sample was recruited like the procedure for new participants for the present recruitment, amongst others through newspaper ads. Finally, confounding will only be present if psychotherapy addresses the biochemical, neuropsychological or brain pathology that is associated with underlying mechanisms. As new insights in underlying pathophysiology point to a role of bias modification (Browning et al. 2012) and cognitive control training (Siegle et al. 2014) as efficacious ingredients in recurrence prevention, it is also questionable whether conventional psychotherapies (and even PCT) despite their effects in recurrence prevention indeed change underlying mechanisms.

VERSION 2 - REVIEW

REVIEWER	Jeremy G. Stewart McLean Hospital, Harvard Medical School United States of America
REVIEW RETURNED	22-Aug-2015

GENERAL COMMENTS	The authors have adequately addressed all my previous concerns from their previous submission. I have no further comments on the revised manuscript.

REVIEWER	Dr Stephen-Mark Cooper Cardiff School of Sport Cardiff Metropolitan University Cyncoed Campus Cyncoed Road

	Cyncoed CARDIFF, CF24 6XD WALES, UK
REVIEW RETURNED	04-Nov-2015

GENERAL COMMENTS	In this manuscript (MS) the authors are proposing a cohort design research protocol to investigate the efficacy of advancing knowledge on four issues: i) factors that are associated with recurrent major depressive disorder (MDD) vulnerability, ii) how these factors co-relate, iii) the predictive relationships of these factors with prospective recurrence, and, iv) the changes observed in these factors during recurrence. As I am not an expert on the body of research that is the focus of this proposed protocol (mental health), consequently I have limited myself to specifically concentrating on the research design and the proposed statistical analyses of the data once collected. In my view, the study is well-designed, the research is well-organised, and it would seem that the data will be collected appropriately. As far as the methodology is concerned, I am convinced that the study could be replicated based upon the description the authors provide in the MS. The authors communicate the necessary information in a clear manner in a MS that is generally well-crafted and skilfully written. I found the protocol interesting, relevant and contemporary however it was overlong. In terms of the statistical plan proposed, and whilst the authors have presented a substantial narrative around their a priori power analysis (p 21 & 22), I still think the MS is a little vague on how the authors actually arrive at their (relatively small) sample sizes of an experimental group of n = 40 recurrent MDD-patients and a control group of n = 40. In addition, I wondered what the justification was for setting a priori alpha at a probability of 10% (P < 0.1) in the paragraph on p 21 dealing with the 1st and 2nd hypotheses (line 30)? It might also be useful for the authors to consider a post-hoc power analysis of their outcomes once the data have been analysed. This MS complies with the STROBE 2007 (v4) criteria for cohort studies.

VERSION 2 – AUTHOR RESPONSE

Reviewer #1

The authors have adequately addressed all my previous concerns from their previous submission. I have no further comments on the revised manuscript.

We would like to again thank Reviewer #1 for his contributions to the manuscript.

Reviewer #2

In this manuscript (MS) the authors are proposing a cohort design research protocol to investigate the efficacy of advancing knowledge on four issues: i) factors that are associated with recurrent major depressive disorder (MDD) vulnerability, ii) how these factors co-relate, iii) the predictive relationships

of these factors with prospective recurrence, and, iv) the changes observed in these factors during recurrence.

As I am not an expert on the body of research that is the focus of this proposed protocol (mental health), consequently I have limited myself to specifically concentrating on the research design and the proposed statistical analyses of the data once collected. In my view, the study is well-designed, the research is well-organised, and it would seem that the data will be collected appropriately.

As far as the methodology is concerned, I am convinced that the study could be replicated based upon the description the authors provide in the MS. The authors communicate the necessary information in a clear manner in a MS that is generally well-crafted and skillfully written. I found the protocol interesting, relevant and contemporary however it was overlong. This MS complies with the STROBE 2007 (v4) criteria for cohort studies.

We would like to thank Reviewer #2 for his positive comments regarding our protocol.

1. In terms of the statistical plan proposed, and whilst the authors have presented a substantial narrative around their a priori power analysis (p 21 & 22), I still think the MS is a little vague on how the authors actually arrive at their (relatively small) sample sizes of an experimental group of n = 40 recurrent MDD-patients and a control group of n = 40.

We agree that despite our relatively elaborate power analyses, the rationale for our final sample size could be better explained. Of note, the sample sizes are for the experimental group **n=60** of recurrent MDD-patients and a control group of n = 40, i.e. 25% larger than the n = 40 vs. n = 40 that was mentioned by Reviewer #2.

Our power analyses show that with these sample sizes we have adequate power to detect small to medium effect sizes. Of course we would have liked to obtain more data, but feasibility aspects limit further increases in sample size. First, MRI scanning costs limit subject number, but also recruitment effort in order to find these specific subjects with remitted recurrent MDD that are medication free need to remain manageable. We now more clearly explained this balance in section 2.5.3.1 of our revised manuscript.

2. In addition, I wondered what the justification was for setting a priori alpha at a probability of 10% (P < 0.1) in the paragraph on p 21 dealing with the 1st and 2nd hypotheses (line 30)?

This P of .1 was chosen to identify potential confounders, i.e. not for hypothesis testing. This is a standard procedure: by setting a more lenient alpha you make sure no potential confounders are overlooked. For testing our hypotheses we will use an alpha of .05 unless otherwise explained, as stated in our power calculation. We now replaced this sentence in section 2.5.3.3 to further clarify this.

3. It might also be useful for the authors to consider a post-hoc power analysis of their outcomes once the data have been analysed.

We thank Reviewer #2 for this suggestion, which we incorporated in our manuscript in section 2.5.3.1.

VERSION 3 – REVIEW

REVIEWER	Dr Stephen-Mark Cooper Cardiff School of Sport Cardiff Metropolitan University Cyncoed Campus Cyncoed Road Cyncoed CARDIFF, CF24 6XD WALES, UK
REVIEW RETURNED	04-Jan-2016

GENERAL COMMENTS	As mentioned in my first review as far as the methodology is concerned, I am convinced that the study could be replicated based upon the description the authors provide in the MS. The information is communicated in a clear manner and the protocol proposed is interesting, relevant and contemporary. However, I still think that the proposal is overlong, but I would not want to hold it up on that fact alone. Provided that it meets with the journal's word limit policy, of course! In terms of the statistical plan proposed, the concerns that I expressed in my first review have been adequately dealt with in this revision of the proposal.