

No Causal Links Between Estradiol and Female's Brain and Mental Health Using Mendelian Randomization

Corresponding Author: Ms Hannah Oppenheimer

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

In this manuscript, Oppenheimer and colleagues present findings from a two-sample Mendelian randomization analysis that examines potential causal relationships between measures of lifetime estradiol exposure and multiple health outcomes (namely brain age, Alzheimer's disease, and depression). The foundational premise of this study is sound, as examination of the relationships among female-specific health factors and later-life brain and health outcomes is without question of great importance. It also appears that their use of the MR method was appropriate. However, I believe that there are issues with the selection of exposure variables that need to be addressed before the study can be considered suitable for publication. These issues bring into question the overall validity of the study design, and whether MR is indeed the best approach to address the questions relevant to this study. I outline my concerns below:

1. The binary estradiol phenotype strikes me as a deeply flawed phenotype. The authors cite studies that have used this phenotype in the past; nevertheless, as they describe to some extent in the discussion, the phenotype is without question confounded by age of participants and will likely be more representative of a classification of menopause status than anything else. Given the inherent flaws of this phenotype, which the authors acknowledge in the Discussion, why is it being used as one of the primary exposure variables?
2. The continuous estradiol measure is far superior phenotype; however, it appears that it was not collected with any consideration of the timing of women's menstrual cycle. Estradiol levels can vary dramatically over the course of a women's cycle, and failing to account for this source of variation would likely introduce significant error into any GWAS analysis. Was cycle phase considered in the data collect? If not, this limitation or caveat needs to be acknowledged somewhere in the manuscript.
3. How do the authors conceptualize the use of genetic prediction of exposures such as history of oral contraception use, hormone replacement therapy, history of hysterectomy, and history of oophorectomy. While I do not question that there is likely some genetic component to these phenotypes, do the authors believe that these genetic factors will have causal relationships with the outcomes? Consideration needs to be given to the question of where it is appropriate to examine the genetic predictors of lifestyle of health decisions in the context of an MR analysis.
4. The potential role of HRT as a risk or protective factor for later-life disease is indeed an important question, but the key issues around HRT are more related to duration of use, formulations, and timing of use (e.g., within a critical window of the menopause transition). The current design is not able to address any of these issues, so I question where or not this variable should be used as an exposure.
5. In the selection of the exposure variables, there appears to be inconsistency in how the mechanism of menopause is considered (i.e., whether it is spontaneous or induced). As described in Table 1, age at natural menopause was used for reproductive span, while age at last period was used for age at menopause. The distinctions are subtle, but they could have a significant impact on the phenotype in question.
6. Regarding the use of history of hysterectomy as an exposure, how relevant is a hysterectomy without oophorectomy to lifetime estrogen exposure? Did the GWAS for this exposure consider hysterectomy conducted after menopause?

In addition to the above notes, I add the following suggestions:

7. Please review the references cited in the second paragraph of the Introduction, specifically those pertaining to studies that have used data from the UK Biobank. I don't think that all of these are associated with brain age, and I'm pretty sure some of the results contradict each other.

8. The authors should clearly describe the assumptions of the MR, and how these assumptions impact the interpretability of the results.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)

This work sets out sex-specific two-sample Mendelian randomisation analysis of the effects of life-long exposure to estradiol on mental health related outcomes, including brain age gap, depression and alzheimer's. This involved performing the first sex-specific brain age gap GWAS in UK Biobank data. The researchers found no causal associations between estradiol levels or other indicators of estradiol exposure through the life course with any of the outcomes. This study contributes to the important field of sex-specific genetic analysis to identify disease aetiologies, but is limited by the lack of appropriate GWAS data on estradiol levels in women.

Strengths:

1. Contributes to the field of sex-specific disease aetiology
2. Sex-specific brain age gap GWAS is a useful contribution to the field
3. Good use of sensitivity analyses to validate fulfilment of MR assumptions

Major comments

1. The use of a binary instrument for estradiol (detectable vs undetectable levels) in UK Biobank raises concerns, as it is more likely representing menopause status in women. The paper considers this in the discussion and highlights the limitations, these concerns are sufficient to potentially undermine the use of these SNPs as instruments. The paper discusses work by Haas et al. 2022 that also uses a binary approach to estradiol GWASs in biobank, however, that paper focussed their results on signals in premenopausal women, to remove the bias introduced by differences in menopausal status. This paper might consider a similar approach. It might also be helpful to explore further the biology that is being detected by the continuous estradiol GWAS hits versus the binary estradiol GWAS hits. Could these SNPs be explored in other data sets where a more accurate measure of estradiol is available, even if the sample is smaller or have less comprehensive genetic measurements? Are there other relevant phenotypes that could be used to provide confidence that these are not just menopause related? If, say, there is a hit that more heavily drives estradiol levels, this SNP could be the focus of the analysis, similarly to the way SNPs in the SHBG gene are centralised in biological studies of SHBG levels.
2. GWASs of history of hysterectomy and oophorectomy are more likely to detect signals associated with cancers of the reproductive system than exposure to estradiol through the life course, but this is not considered in the paper. It may be better to exclude these exposures for brevity and clarity, especially given that at the standard GWAS significance threshold only 3 SNPs were identified for history of hysterectomy, and 6 for oophorectomy.
3. The regional plots of genomic loci identified in estradiol levels (in the continuous approach) (supplementary figure 7) are suggestive of noise rather than true signals, as the lead SNP in many cases appears to be independent of other SNPs at the same locus. This suggests there may be problems with this instrument that require further investigation before drawing conclusions about the causal association between continuous estradiol levels and mental health conditions.
4. It would be useful to the reader if the paper outlined more specifically how proxies were identified, particularly (if it were possible) with this added to the associated code files. The paper describes in the supplementary material that the proxy variants were searched for using the LDProxy Tool from LDlink, but a more detailed explanation of how this identifies proxies would benefit readers, especially given the large discrepancies in SNPs identified in the depression and alzheimer's datasets when the same exposures are being used.
5. The paper is very long, with an especially long discussion with subheadings.

Minor comments

1. Figure 1a(3) is a low quality table which cannot be read when zoomed in.
2. Table 1, by crossing two pages and including a large amount of information is difficult to read and may be more effective in the supplement with landscape formatting. It would also be useful (given the differences in numbers of SNPs used) to show what the source of the data for the different instruments was by rows.
3. The paper could attempt to explain their findings further in the discussion. There is lots of description of the findings. For example, the attenuation of the significant causal association between age at menarche and depression once BMI was included is described, but an explanation as to why this might be would benefit readers seeking to make biological sense of Mendelian randomisation analyses.
4. It would seem prudent to discuss fully, when covering the limitations, that for many of the MR analyses, the small number of SNPs means that the power of these analyses would be quite low.

(Remarks on code availability)

Code appears to be standard for this type of analysis, but it would be useful to have the code that was used to identify any

proxies used in this analysis.

Reviewer #3

(Remarks to the Author)

This was an interesting and well written study by Oppenheimer et al looking at estradiol exposure and the female brain and mental health. The main result is that estradiol exposure is not causing depression or altering the female brain age gap as might be expected based on the observational literature. The methodology used is generally sound although I do have a few questions:

1. Is UKB the best option for a GWAS of estradiol levels? Given the age range of the group many women have gone through menopause and this means that most women in the study have a reading outside the detectable limit. Does the binary metric you have used effectively capture menopause rather than hormone levels? This was previously discussed in Ruth et al: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7025895/> I'd certainly question how accurate the exposure was at solely capturing estradiol levels. For comparison could you look at estradiol levels in men, which did seem to somewhat heritable in men?
2. When you only have 8 or 9 exposure loci how reliable is MR Egger? I believe that when 10 or less instruments it is a fairly imprecise tool.
3. What do the betas in table 2 reflect for estradiol? Is it based on the binary metric? Therefore should convert to a doubling in genetic liability for a binary exposure? <https://pubmed.ncbi.nlm.nih.gov/30039250/>
4. Given low numbers of exposure in some analysis and the issues with overlap which might limit sample size worth considering using MRlap as a method?
5. Did you look at the potential pleiotropic nature of the variants included as exposure? For example, did you look them up in the GWAS catalogue.

(Remarks on code availability)

Code is clear and easy to follow. Well annotated and from a MR perspective looks sensible.

Reviewer #4

(Remarks to the Author)

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

(Remarks on code availability)

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

In this revised manuscript the authors have been very responsive to this reviewer's prior comments. While I still have questions regarding whether the MR approach is the most appropriate technique for examining the questions addressed in this study, the manuscript is now appropriate for publication in my opinion. I hope that this study will stimulate a healthy debate within the field.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)

It seems that there have been some substantial changes to the underlying instrument used in this study, and I think that that puts in on a much more secure footing that avoids the previous issues of confounding.

However, now that the main analysis is the 2 loci from the continuous exposure, I wonder to what extent this can be described as "life-time exposure", given that most of the sample is post-menopausal.

The other effect is that now the results section is quite confusing, with a number of different scores being proposed and reported in table 1 - there are now three possible samples (pre, post and combined) and two possible measures (continuous and binary) and that's just in UK Biobank, with the LIFE samples as a addition. This made it hard to track the numbers of signals through the paper - especially as the authors have reported some loci as genome-wide significant, mentioned multiple signals at these loci, but then used in some cases a lower threshold to determine instrument selection. It would be useful to have a table somewhere where just the GWAS number of GWAS signals were reported, and separately the number of instruments that were used. I'd also be tempted to suggest splitting Table 1 into one table with just the Estradiol results. And the other phenotypes separately, as more detail about the Estradiol instruments could then be given. Also, while a lower

threshold is mentioned in the methods section, I think that this could have been briefly mentioned in the results section to make the article easier to follow.

I thank the authors for the clarification of the method to find proxies, but it still seems a bit surprising that the SNP numbers are so low for Depression compared to the other phenotypes.

With regard to the limitations of UKB, a major issue is the overall rate of detection of the hormone, while this is mentioned on line 348, it would be useful if the authors could discuss how they think it would have specifically impacted their study.

Feedback on the responses to reviewer 3's previous comments:

The authors responses to points 1-3 are all comprehensive and sufficient.

Point 4. Regarding the use of MRlap to address any possible issues of sample overlap, it appears from Supplementary Table 8 that MRlap was only used for a subset of the analyses. The authors correctly suggest that MRlap is only of benefit where there is the possibility of sample overlap. However, for a number of the traits that are considered in this paper, there is more than one data set used (as described in Table 2). And for some of the phenotypes listed in Supplementary Table 8, there could be sample overlap depending on the data set chosen for the analysis. I'm aware that the authors have made some effort to try to have independent samples where this was possible, but to make this clear to the reader, could a signifier (probably first author and year) be used to indicate the source in Supplementary Table 8. It might also be helpful to do this for age at menarche and age at menopause in Table 1 in the main text.

Point 5. While this data is very comprehensive, the formatting of Table 14 and 15 could be better. It's not clear what is different between the columns "Independent Significant SNP" and "SNP". The caption should explain this more clearly. There are also multiple rows where information is exactly duplicated in other rows. I'm also not sure that the added sentence on pages 23/24 is quite sufficient. The authors spend some time in the discussion talking about the overlap with SHBG, so it would make sense to also mention it here.

(Remarks on code availability)

No further comments on the code after last time.

Reviewer #4

(Remarks to the Author)

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

(Remarks on code availability)

Version 2:

Reviewer comments:

Reviewer #2

(Remarks to the Author)

I thank the authors for the changes that they have made, and have no further comments.

(Remarks on code availability)

Reviewer #4

(Remarks to the Author)

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

(Remarks on code availability)

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

In this manuscript, Oppenheimer and colleagues present findings from a two-sample Mendelian randomization analysis that examines potential causal relationships between measures of lifetime estradiol exposure and multiple health outcomes (namely brain age, Alzheimer's disease, and depression). The foundational premise of this study is sound, as examination of the relationships among female-specific health factors and later-life brain and health outcomes is without question of great importance. It also appears that their use of the MR method was appropriate. However, I believe that there are issues with the selection of exposure variables that need to be addressed before the study can be considered suitable for publication. These issues bring into question the overall validity of the study design, and whether MR is indeed the best approach to address the questions relevant to this study. I outline my concerns below:

General response: We thank the reviewer for taking the time to evaluate our manuscript, and for the helpful suggestions for improvements. Please find our responses to each comment below, and the corresponding edits highlighted in yellow in the uploaded copy of the revised manuscript file.

1. The binary estradiol phenotype strikes me as a deeply flawed phenotype. The authors cite studies that have used this phenotype in the past; nevertheless, as they describe to some extent in the discussion, the phenotype is without question confounded by age of participants and will likely be more representative of a classification of menopause status than anything else. Given the inherent flaws of this phenotype, which the authors acknowledge in the Discussion, why is it being used as one of the primary exposure variables?

Response: We thank the reviewer for this comment and agree that the binary estradiol phenotype is likely confounded by the age of the participants. Following this comment, we have replaced the binary estradiol phenotype with the continuous estradiol phenotype in the main manuscript and moved the analyses using the binary estradiol phenotype into the supplements.

Moreover, we have replicated the results on continuous estradiol levels in additional samples using openly available GWAS summary statistics. We have now included an independent sample of postmenopausal females from the LIFE-Adult and LIFE-Heart study (Pott et al., 2019), as well as a sample of only premenopausal females and a sample of only postmenopausal females from the UK Biobank, using both the continuous and the binary estradiol phenotypes (Haas et al., 2022). Following suggestions in comment 1 of reviewer #3, we have additionally included analyses in male-only samples using continuous estradiol levels measured in the UK Biobank (<http://www.nealelab.is/uk-biobank>). Across all samples, we consistently did not find any significant associations between estradiol levels and brain age gap, Alzheimer's disease, and depression. The results remained robust throughout sensitivity analyses using the binary estradiol phenotype.

We have now included these additional analyses in the methods section of the revised manuscript under *Samples and Data Sources: Available Summary Statistics* (page 17) as follows:

“To ensure robustness to effects of menopausal status previously shown to impact GWAS results²⁴, several datasets were included for estradiol levels. The following samples from the UKB were included: a combined sample of pre- and postmenopausal females, a sample of only premenopausal females, and a sample of only postmenopausal females²⁴. Main analyses in each sample were performed using GWAS of continuous estradiol levels. Sensitivity analyses were conducted in each sample using a binary approach examining estradiol levels above and below the detection limit²⁴ to increase sample sizes. Further, analyses on continuous estradiol levels were replicated in an independent sample of postmenopausal females from the LIFE-Adult & LIFE-Heart studies⁴⁵ and a male-only sample from the UKB⁴⁶.”

Further, we have included these datasets and analyses in Table 2 (page 20-21) and Supplementary Table 1 (Supplementary Information page 28-29). The results of the analyses are mentioned in the revised manuscript under *Mendelian Randomization Analyses* (page 7) as follows:

“No significant causal relationships were found between estradiol levels, using the continuous measures as the exposure, and brain age gap, Alzheimer's disease, and depression as outcomes (Table 1). The results were consistently non-significant across the combined, premenopausal-only, and postmenopausal-only samples from the UKB. Similarly, the replication analyses in the independent postmenopausal sample from the LIFE studies were non-significant (Table 1).

All results remained non-significant across the estimation methods, except for the analysis with estradiol levels in the combined pre- and postmenopausal sample as an exposure and Alzheimer's disease as an outcome, which was significant using the MR-Egger method ($b = 1.27$, $se = 0.48$, $p = .02$; Supplementary Table 8). Further, the sensitivity analyses using binary estradiol levels for the combined pre- and postmenopausal, premenopausal-only, and postmenopausal-only females from the UKB were not significant for any of the outcomes (Supplementary Table 9). For the male samples, no significant associations were found between estradiol levels and brain age gap, Alzheimer's disease, and depression (Supplementary Table 9). The results remained non-significant across the estimation methods (Supplementary Table 8)."

2. The continuous estradiol measure is far superior phenotype; however, it appears that it was not collected with any consideration of the timing of women's menstrual cycle. Estradiol levels can vary dramatically over the course of a women's cycle, and failing to account for this source of variation would likely introduce significant error into any GWAS analysis. Was cycle phase considered in the data collect? If not, this limitation or caveat needs to be acknowledged somewhere in the manuscript.

Response: We thank the reviewer for this comment and agree that menstrual cycle dependent variations in estradiol levels are important to consider. As we include both premenopausal and postmenopausal females in our GWAS on estradiol levels, we did not account for cycle phase in the present study. However, we have now included analyses using GWAS conducted in a sample of only premenopausal females from the UK Biobank (Haas et al., 2022). In this GWAS, Haas et al. (2022), conducted sensitivity analyses adjusting for menstrual cycle timing and did not find any additional signals. Further, we have included analyses using a GWAS of estradiol levels conducted in a sample of only postmenopausal females, which was adjusted for time since last period (Haas et al., 2022).

Further, in order to limit biases in our own GWAS, we have rerun our GWAS on continuous estradiol levels and have now included the following covariates known to influence estradiol levels: menopausal status (premenopausal/postmenopausal), ever taken oral contraceptives (yes/no), ever used hormone replacement therapy (yes/no), history of bilateral oophorectomy (yes/no), and history of hysterectomy (yes/no).

We have described this in the revised manuscript in the methods section under *GWAS Procedure* (page 23) as follows:

“The GWAS on continuous estradiol levels was further controlled for menopausal status (premenopausal/postmenopausal), history of oral contraceptive use (yes/no), history of HRT use (yes/no), history of bilateral oophorectomy (yes/no), and history of hysterectomy (yes/no), as these variables are known to influence estradiol levels.”

And under *Samples and Data Sources: UKB Samples* (page 18) as follows:

“Further, participants who answered, “prefer not to answer” or “do not know/not sure” on any of the covariates included in the GWAS on continuous estradiol levels were excluded (Supplementary Note 1).”

And we have specified this further in Supplementary Note 1 (Supplementary Information page 4-5) as follows:

“Of the $n = 42,084$ females with estradiol levels above the detection limit, $n = 7,387$ females were removed due to answering “prefer not to answer” or “do not know/not sure” on one of the covariates (i.e., menopausal status (“had menopause”; UKB Field 2724), history of oral contraceptive use (“ever taken oral contraceptive pill”; UKB Field 2784), history of HRT use (“ever used HRT”; UKB Field 2814), history of bilateral oophorectomy (“bilateral oophorectomy (both ovaries removed)”; UKB Field 2834), and history of hysterectomy (“ever had hysterectomy (womb removed)”; UKB Field 3591)). This resulted in a final sample with continuous estradiol levels of $N = 34,697$ females.”

Further, we have acknowledged limitations arising from covariates in the revised manuscript in the discussion (page 13) as follows:

“The GWAS of the present study might be limited by the included covariates. For instance, while we controlled for menopausal status as well as history of oral contraceptive use, HRT use, hysterectomy, and oophorectomy, menstrual cycle phase was not considered due to the inclusion of pre- and postmenopausal females, potentially influencing the findings.”

3. How do the authors conceptualize the use of genetic prediction of exposures such as history of oral contraception use, hormone replacement therapy, history of hysterectomy, and history of oophorectomy. While I do not question that there is likely some genetic component to these phenotypes, do the authors believe that these genetic factors will have

causal relationships with the outcomes? Consideration needs to be given to the question of where it is appropriate to examine the genetic predictors of lifestyle of health decisions in the context of an MR analysis.

Response: We thank the reviewer for this comment and agree that this is an important point to consider. There are not many previous studies examining these factors using genetic approaches. However, one study using drug-target Mendelian randomization found a significant causal association between genetically predicted perturbation of estrogen receptor (ER) β and higher depression, pointing to implications for drugs such as hormone replacement therapy and oral contraceptives (Schindler et al., 2024). While hormone replacement therapy use, oral contraceptive use, history of hysterectomy, and history of oophorectomy are lifestyle health decisions or health-related procedures, the use of these variables in the present study contributes to the field examining the genetic contribution, albeit it may be secondary to environmental influences. Nevertheless, in line with this comment as well as comments 4 and 6, we have decided to move these analyses to the supplements and mention their limitations in the discussion.

We mention these supplementary analyses in the methods section of the revised manuscript under *Samples and Data Sources: Available Summary Statistics* (page 17) as follows:

“Supplementary analyses were conducted using factors related to exogenous hormone use and health-related procedures likely impacting estradiol levels, including oral contraceptive use, HRT use, history of hysterectomy, and history of oophorectomy⁵⁰ as exposure variables.”

And in the results section under *Mendelian Randomization Analyses* (page 8):

“Consistently, no significant effects were found in the supplementary analyses using oral contraceptive use, HRT use, history of hysterectomy, and history of oophorectomy as exposure variables (Supplementary Note 5 and Supplementary Table 7).”

Further, we discuss this limitation in the discussion of the revised manuscript (page 13-14):

“However, it must be mentioned that some of the additionally analyzed exposure variables, such as HRT use and oral contraceptive use, might not have a substantial genetic component, potentially leading to null findings in the present study. Furthermore, non-linear and age-dependent effects, as have been suggested for associations of number of childbirths and HRT

use with brain health^{7,36}, should be investigated. For instance, previous studies highlight the importance of timing, duration, and formulation of HRT use³⁶. These aspects were not captured by the variable used for HRT use in the present study, possibly resulting in null findings. Further, the GWAS of history of hysterectomy did not take important influencing factors into account, such as the medical reason for the procedure, the timing of procedure (i.e., before or after menopause), and whether an oophorectomy also took place. These limitations can be traced back to the use of GWAS from consortia, which have advanced research through publicly available, large-scale datasets, however, are often designed in a standardized manner across sexes and may not consider potentially relevant covariates for specific variables or samples. Future research is needed to disentangle genetic and environmental contributions to these variables and further assess their causal associations with brain and mental health.”

4. The potential role of HRT as a risk or protective factor for later-life disease is indeed an important question, but the key issues around HRT are more related to duration of use, formulations, and timing of use (e.g., within a critical window of the menopause transition). The current design is not able to address any of these issues, so I question where or not this variable should be used as an exposure.

Response: We thank the reviewer for this comment. As mentioned in our response to comment 3, we now discuss this limitation in the discussion of the revised manuscript (page 13-14):

“However, it must be mentioned that some of the additionally analyzed exposure variables, such as HRT use and oral contraceptive use, might not have a substantial genetic component, potentially leading to null findings in the present study. Furthermore, non-linear and age-dependent effects, as have been suggested for associations of number of childbirths and HRT use with brain health^{7,36}, should be investigated. For instance, previous studies highlight the importance of timing, duration, and formulation of HRT use³⁶. These aspects were not captured by the variable used for HRT use in the present study, possibly resulting in null findings.”

5. In the selection of the exposure variables, there appears to be inconsistency in how the mechanism of menopause is considered (i.e., whether it is spontaneous or induced). As described in Table 1, age at natural menopause was used for reproductive span, while age at last period was used for age at menopause. The distinctions are subtle, but they could have a significant impact on the phenotype in question.

Response: We thank the reviewer for this comment and for making us aware about the different mechanisms of menopause included in the exposure variables of age at menopause and reproductive span. The exposure variable measuring age at menopause as the last menstrual period (Loh et al., 2018) was chosen due to the larger sample size. To address the concerns of the different mechanisms, we have now added sensitivity analyses using age at natural menopause (Day et al., 2015) in addition to age at menopause (last menstrual period; Loh et al., 2018). The results remained consistent throughout analyses.

We mention the sensitivity analyses in the methods under *Samples and Data Sources: Available Summary Statistics* (page 17) of the revised manuscript as follows:

“As the GWAS on age at menopause used for Alzheimer’s disease and depression as outcomes was not limited to natural menopause⁴⁷, sensitivity analyses were conducted using age at natural menopause⁴⁸. To avoid sample overlap, analyses using age at menarche⁴⁹ and age at natural menopause⁴⁸ as exposures and brain age gap as an outcome were conducted using GWAS summary statistics from a smaller, independent sample.”

And in the results section (page 8):

“Similarly, age at natural (non-surgical) menopause as an exposure was not significant in any of the sensitivity analyses (Supplementary Table 10).”

6. Regarding the use of history of hysterectomy as an exposure, how relevant is a hysterectomy without oophorectomy to lifetime estrogen exposure? Did the GWAS for this exposure consider hysterectomy conducted after menopause?

Response: We thank the reviewer for this comment and acknowledge that a history of oophorectomy in the context of a hysterectomy and that the timing of a hysterectomy (i.e., before or after menopause) are highly relevant. However, hysterectomies alone have also been shown to impact ovarian function, increase risk of ovarian failure, and potentially lead to an earlier menopause (Huang et al., 2023; Moorman et al., 2011). Further, both hysterectomies with or without oophorectomy, compared to not having had a hysterectomy, have been found to increase risk of dementia in observational studies (Gilsanz et al., 2019; Gong et al., 2022).

The GWAS used for the exposure of history of hysterectomy did not consider whether an oophorectomy was performed and the timing of the hysterectomy relative to menopause. The GWAS was based on all hysterectomies included in the UK Biobank Field 3591 (Elsworth

et al., 2018). We were unfortunately not able to find a previously conducted GWAS that considered these important points. Therefore, following this comment and comment 3, we have decided to move these analyses to the supplements.

We mention these supplementary analyses in the methods section of the revised manuscript under *Samples and Data Sources: Available Summary Statistics* (page 17) as follows:

“Supplementary analyses were conducted using factors related to exogenous hormone use and health-related procedures likely impacting estradiol levels, including oral contraceptive use, HRT use, history of hysterectomy, and history of oophorectomy⁵⁰ as exposure variables.”

And in the results section under *Mendelian Randomization Analyses* (page 8):

“Consistently, no significant effects were found in the supplementary analyses using oral contraceptive use, HRT use, history of hysterectomy, and history of oophorectomy as exposure variables (Supplementary Note 5 and Supplementary Table 7).”

Further, we discuss this limitation in the discussion of the revised manuscript (page 13-14):

“However, it must be mentioned that some of the additionally analyzed exposure variables, such as HRT use and oral contraceptive use, might not have a substantial genetic component, potentially leading to null findings in the present study. Furthermore, non-linear and age-dependent effects, as have been suggested for associations of number of childbirths and HRT use with brain health^{7,36}, should be investigated. For instance, previous studies highlight the importance of timing, duration, and formulation of HRT use³⁶. These aspects were not captured by the variable used for HRT use in the present study, possibly resulting in null findings. Further, the GWAS of history of hysterectomy did not take important influencing factors into account, such as the medical reason for the procedure, the timing of procedure (i.e., before or after menopause), and whether an oophorectomy also took place. These limitations can be traced back to the use of GWAS from consortia, which have advanced research through publicly available, large-scale datasets, however, are often designed in a standardized manner across sexes and may not consider potentially relevant covariates for specific variables or samples. Future research is needed to disentangle genetic and environmental contributions to these variables and further assess their causal associations with brain and mental health.”

In addition to the above notes, I add the following suggestions:

7. Please review the references cited in the second paragraph of the Introduction, specifically those pertaining to studies that have used data from the UK Biobank. I don't think that all of these are associated with brain age, and I'm pretty sure some of the results contradict each other.

Response: We agree that this paragraph was misleading and have rephrased it in the introduction of the revised manuscript (page 3) to better reflect the mentioned findings:

“Further, these factors have been associated with markers of brain health using data from the UK Biobank (UKB)⁷⁻⁹, a prospective population-based cohort from the United Kingdom (see Sudlow et al., 2015¹⁰ for details on the UKB). For instance, lower brain age gap has been linked to a higher number of childbirths⁷, an older age at natural menopause⁸, and a longer reproductive span⁹.”

8. The authors should clearly describe the assumptions of the MR, and how these assumptions impact the interpretability of the results.

Response: We thank the reviewer for this comment and we agree that this is a relevant point. Therefore, we have added the following information to the methods section under *Two-Sample Mendelian Randomization Analyses: Selection of Instrumental Variables* (page 24) of the revised manuscript:

“The instrumental variables included in a Mendelian randomization analysis are assumed to be associated with the respective exposure variable (relevance assumption), not be associated with the outcome variable via confounding pathways (exchangeability assumption), and not affect the outcome variable in any way other than through the exposure variable (exclusion restriction assumption)¹⁵. The relevance assumption can be verified, while the exchangeability and exclusion restriction assumptions are not verifiable¹⁵, however, a set of sensitivity analyses, including the use of robust methods and multivariable Mendelian randomization, were conducted to assess biases arising from potential violations.

To satisfy the relevance assumption, instrumental variables were selected by applying a genome-wide significance threshold of $p < 5 \times 10^{-8}$ to identify SNPs associated with the respective exposure variables.”

Further, we have expanded the discussion (page 14) of the revised manuscript to include the following:

“Furthermore, it is important to note that while we set a strict threshold for selecting instrumental variables in the Mendelian randomization analyses to satisfy the relevance assumption, the exchangeability and exclusion restriction assumptions cannot be verified and may limit the validity of the results¹⁵. Although we conducted a variety of sensitivity analyses and reduced sample overlap, pleiotropic effects and confounding of the genetic variant-outcome association remain possible, for example arising from population stratification or assortative mating¹⁵.”

Reviewer #2 (Remarks to the Author):

This work sets out sex-specific two-sample Mendelian randomisation analysis of the effects of life-long exposure to estradiol on mental health related outcomes, including brain age gap, depression and alzheimer's. This involved performing the first sex-specific brain age gap GWAS in UK Biobank data. The researchers found no causal associations between estradiol levels or other indicators of estradiol exposure through the life course with any of the outcomes. This study contributes to the important field of sex-specific genetic analysis to identify disease aetiologies, but is limited by the lack of appropriate GWAS data on estradiol levels in women.

Strengths:

- 1. Contributes to the field of sex-specific disease aetiology**
- 2. Sex-specific brain age gap GWAS is a useful contribution to the field**
- 3. Good use of sensitivity analyses to validate fulfilment of MR assumptions**

General response: We thank the reviewer for taking the time to evaluate our manuscript, and for the helpful suggestions for improvements. Please find our responses to each comment below, and the corresponding edits highlighted in yellow in the uploaded copy of the revised manuscript file.

Major comments

1. The use of a binary instrument for estradiol (detectable vs undetectable levels) in UK Biobank raises concerns, as it is more likely representing menopause status in women. The paper considers this in the discussion and highlights the limitations, these concerns are sufficient to potentially undermine the use of these SNPs as instruments. The paper discusses work by Haas et al. 2022 that also uses a binary approach to estradiol GWASs in biobank, however, that paper focussed their results on signals in premenopausal women, to remove the bias introduced by differences in menopausal status. This paper might consider a similar approach. It might also be helpful to explore further the biology that is being detected by the continuous estradiol GWAS hits versus the binary estradiol GWAS hits. Could these SNPs be explored in other data sets where a more accurate measure of estradiol is available, even if the sample is smaller or have less comprehensive genetic measurements? Are there other relevant phenotypes that could be used to provide

confidence that these are not just menopause related? If, say, there is a hit that more heavily drives estradiol levels, this SNP could be the focus of the analysis, similarly to the way SNPs in the SHBG gene are centralised in biological studies of SHBG levels.

Response: We thank the reviewer for this comment and agree that the binary estradiol phenotype may be flawed and confounded by the menopausal status of the participants. Following this comment, we have included additional samples for our analyses on estradiol levels, including the mentioned GWAS by Haas et al. (2022). We have now included separate samples of premenopausal and postmenopausal females, each using both the continuous and binary approaches (Haas et al., 2022). In line with suggestions made in comment 1 by reviewer #1, we have moved the binary phenotypes into the supplements and have included the continuous phenotypes in the main manuscript instead.

Further, we have replicated the analyses in a separate sample of females from the LIFE-Adult and LIFE-Heart study (Pott et al., 2019) and following suggestions in comment 1 of reviewer #3, we have also replicated the analyses in a male sample from the UK Biobank (<http://www.nealelab.is/uk-biobank>).

We have included these additional analyses in the methods section of the revised manuscript under *Samples and Data Sources: Available Summary Statistics* (page 17) as follows:

“To ensure robustness to effects of menopausal status previously shown to impact GWAS results²⁴, several datasets were included for estradiol levels. The following samples from the UKB were included: a combined sample of pre- and postmenopausal females, a sample of only premenopausal females, and a sample of only postmenopausal females²⁴. Main analyses in each sample were performed using GWAS of continuous estradiol levels. Sensitivity analyses were conducted in each sample using a binary approach examining estradiol levels above and below the detection limit²⁴ to increase sample sizes. Further, analyses on continuous estradiol levels were replicated in an independent sample of postmenopausal females from the LIFE-Adult & LIFE-Heart studies⁴⁵ and a male-only sample from the UKB⁴⁶.”

Further, we have included these datasets and analyses in Table 2 (page 20-21) and Supplementary Table 1 (Supplementary Information page 28-29). The results of the analyses are mentioned in the revised manuscript under *Mendelian Randomization Analyses* (page 7) as follows:

“No significant causal relationships were found between estradiol levels, using the continuous measures as the exposure, and brain age gap, Alzheimer’s disease, and depression as outcomes (Table 1). The results were consistently non-significant across the combined, premenopausal-only, and postmenopausal-only samples from the UKB. Similarly, the replication analyses in the independent postmenopausal sample from the LIFE studies were non-significant (Table 1). All results remained non-significant across the estimation methods, except for the analysis with estradiol levels in the combined pre- and postmenopausal sample as an exposure and Alzheimer’s disease as an outcome, which was significant using the MR-Egger method ($b = 1.27$, $se = 0.48$, $p = .02$; Supplementary Table 8). Further, the sensitivity analyses using binary estradiol levels for the combined pre- and postmenopausal, premenopausal-only, and postmenopausal-only females from the UKB were not significant for any of the outcomes (Supplementary Table 9). For the male samples, no significant associations were found between estradiol levels and brain age gap, Alzheimer’s disease, and depression (Supplementary Table 9). The results remained non-significant across the estimation methods (Supplementary Table 8).”

We hope to see an increase in large-scale datasets with reliable estradiol measurements in wider age ranges in the future. We mention these limitations and efforts in the discussion (page 13) of the revised manuscript as follows:

“As a sensitivity analysis, we conducted a GWAS on binary estradiol levels (below or above detection limit), which had a substantially larger sample size. Similar to Haas et al., 2022²⁴, we identified different significant loci in the binary estradiol GWAS, suggesting that different traits might be measured using these different approaches. Issues concerning estradiol measurements in the UKB have been previously discussed, including a potential bias towards the detection of loci associated with menopause due to the age of the participants and the substantial number of measurements below the detection limit^{23,24,33}. Many large-scale databases, including the UKB, are conducted on aging cohorts and are confounded by survivor and healthy volunteer biases³⁴. Therefore, we conducted sensitivity analyses across different samples and approaches. Nevertheless, the present study is limited by the available data on estradiol levels and further highlights the need for large-scale, precise measurements conducted in diverse samples and age groups under consideration of female-specific variables – a data gap which has been repeatedly identified³⁵.”

2. GWASs of history of hysterectomy and oophorectomy are more likely to detect signals associated with cancers of the reproductive system than exposure to estradiol through the life course, but this is not considered in the paper. It may be better to exclude these exposures for brevity and clarity, especially given that at the standard GWAS significance threshold only 3 SNPs were identified for history of hysterectomy, and 6 for oophorectomy.

Response: We thank the reviewer for this comment and agree that the phenotypes history of hysterectomy and history of oophorectomy might not be specific in the context of the present analyses. However, hysterectomies and oophorectomies both influence estradiol levels (Faubion et al., 2015; Huang et al., 2023; Moorman et al., 2011) and have been found to increase risk of dementia in observational studies (Gilsanz et al., 2019; Gong et al., 2022). Due to their links to the variables included in the present study, we chose to include these exposures. Nevertheless, following this comment as well as comment 3 by reviewer #1, we have decided to move the analyses with history of hysterectomy and history of oophorectomy, along with the exposures oral contraceptive use and hormone replacement therapy use, to the supplements for brevity and clarity of the main manuscript.

We mention these supplementary analyses in the methods section of the revised manuscript under *Samples and Data Sources: Available Summary Statistics* (page 17) as follows:

“Supplementary analyses were conducted using factors related to exogenous hormone use and health-related procedures likely impacting estradiol levels, including oral contraceptive use, HRT use, history of hysterectomy, and history of oophorectomy⁵⁰ as exposure variables.”

And in the results section under *Mendelian Randomization Analyses* (page 8):

“Consistently, no significant effects were found in the supplementary analyses using oral contraceptive use, HRT use, history of hysterectomy, and history of oophorectomy as exposure variables (Supplementary Note 5 and Supplementary Table 7).”

3. The regional plots of genomic loci identified in estradiol levels (in the continuous approach) (supplementary figure 7) are suggestive of noise rather than true signals, as the lead SNP in many cases appears to be independent of other SNPs at the same locus. This suggests there may be problems with this instrument that require further

investigation before drawing conclusions about the causal association between continuous estradiol levels and mental health conditions.

Response: We thank the reviewer for this comment and for drawing our attention to the noise that was present in the continuous estradiol GWAS. In the revised manuscript, we have now rerun this GWAS. As the data on estradiol levels was positively skewed, we have now performed an inverse rank normalization of the estradiol measurements before running the GWAS. Further, we have now included the following covariates known to influence estradiol levels: menopausal status (premenopausal/postmenopausal), ever taken oral contraceptives (yes/no), ever used hormone replacement therapy (yes/no), history of bilateral oophorectomy (yes/no), and history of hysterectomy (yes/no). With these steps, we were able to reduce the noise in the continuous estradiol GWAS, which we have now included in the main manuscript. We have rerun the Mendelian randomization analyses using the newly run GWAS.

We have described this in the revised manuscript in the methods section under *Estradiol Levels* (page 22-23) as follows:

“As the estradiol levels were positively skewed, we performed an inverse rank normalization on the data.”

And in the methods section under *GWAS Procedure* (page 23) as follows:

“The GWAS on continuous estradiol levels was further controlled for menopausal status (premenopausal/postmenopausal), history of oral contraceptive use (yes/no), history of HRT use (yes/no), history of bilateral oophorectomy (yes/no), and history of hysterectomy (yes/no), as these variables are known to influence estradiol levels.”

And under *Samples and Data Sources: UKB Samples* (page 18) as follows:

“Further, participants who answered, “prefer not to answer” or “do not know/not sure” on any of the covariates included in the GWAS on continuous estradiol levels were excluded (Supplementary Note 1).”

And we have specified this further in Supplementary Note 1 (Supplementary Information page 4-5) as follows:

“Of the $n = 42,084$ females with estradiol levels above the detection limit, $n = 7,387$ females were removed due to answering “prefer not to answer” or “do not know/not sure” on one of

the covariates (i.e., menopausal status (“had menopause”; UKB Field 2724), history of oral contraceptive use (“ever taken oral contraceptive pill”; UKB Field 2784), history of HRT use (“ever used HRT”; UKB Field 2814), history of bilateral oophorectomy (“bilateral oophorectomy (both ovaries removed)”; UKB Field 2834), and history of hysterectomy (“ever had hysterectomy (womb removed)”; UKB Field 3591)). This resulted in a final sample with continuous estradiol levels of $N = 34,697$ females.”

We report the results of the newly run GWAS on continuous estradiol levels in the revised manuscript under *Results: GWAS of Estradiol Levels in Females* (page 6-7) as follows:

“For the GWAS conducted on continuous estradiol levels in a combined sample of pre- and postmenopausal females, we identified two independent genomic loci, on chromosomes 17 and 19, that were significantly associated with estradiol levels (see Figure 2 for Manhattan plot and Supplementary Figure 4 for QQ-plot). The loci included 3 lead SNPs and 4 independent SNPs and mapped onto 21 genes (Supplementary Table 5 and Supplementary Figure 8). A cluster of genes (SHBG and SLC35G6) expressed lower, on average, across tissue types, compared to the other genes (Supplementary Figure 11). The gene-set enrichment analysis identified 1 positional gene set and 10 associated phenotypes (Supplementary Figure 14). We identified nine independent genomic loci in the GWAS conducted on binary estradiol levels (see Supplementary Note 4 for details).”

While the original GWAS on continuous estradiol levels had identified 33 significant independent genomic loci across different chromosomes, we identified 2 in the newly run GWAS. The Manhattan plot of the newly run GWAS on continuous estradiol levels (page 7) and the regional plots of genomic loci (Supplementary Information, page 19) show reduced noise:

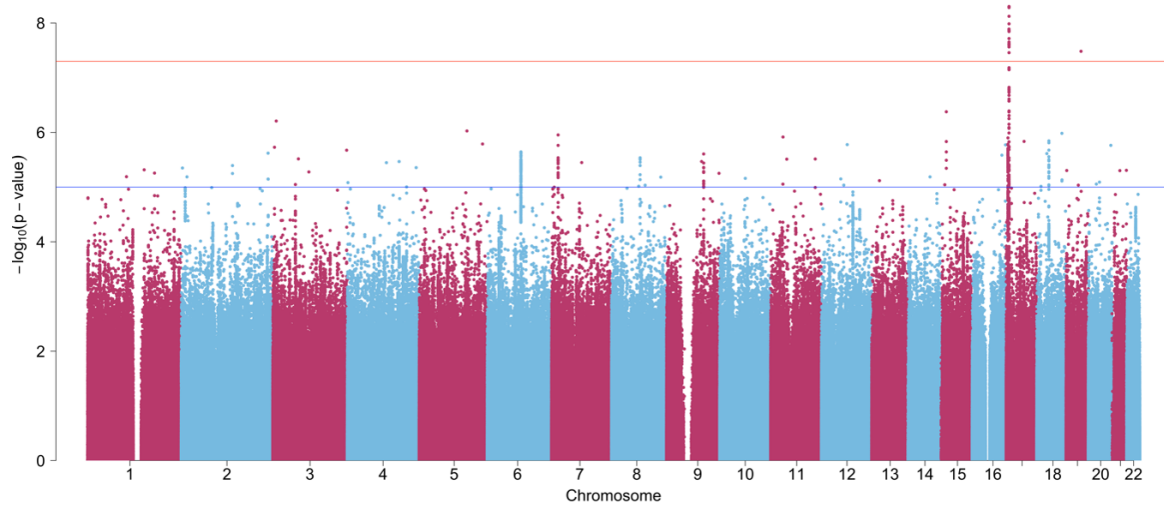


Figure 2. Manhattan Plot of Estradiol Levels (Continuous Approach). The blue and red lines indicate the suggestive and the genome-wide significance thresholds, respectively.

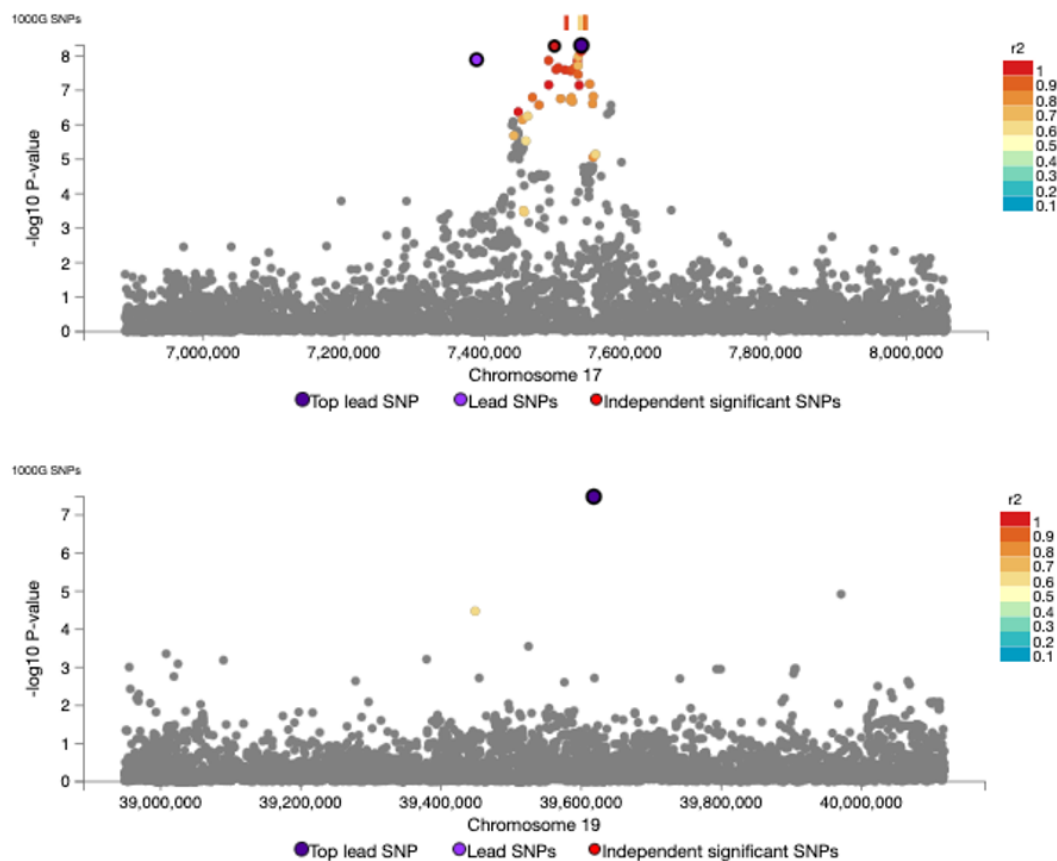


Figure 8. Regional Plots of Genomic Loci identified in Estradiol Levels (Continuous Approach). Plot produced using FUMA, showing top lead SNP, lead SNPs, and independent significant SNPs in each of the identified genomic loci⁶. SNPs are color-coded according to their R^2 values.

The results of the univariable Mendelian randomization analyses using the new GWAS summary statistics are included in the results section of the revised manuscript under *Mendelian Randomization Analyses* (page 7) as follows:

“No significant causal relationships were found between estradiol levels, using the continuous measures as the exposure, and brain age gap, Alzheimer’s disease, and depression as outcomes (Table 1). The results were consistently non-significant across the combined, premenopausal-only, and postmenopausal-only samples from the UKB. Similarly, the replication analyses in the independent postmenopausal sample from the LIFE studies were non-significant (Table 1). All results remained non-significant across the estimation methods, except for the analysis with estradiol levels in the combined pre- and postmenopausal sample as an exposure and Alzheimer’s disease as an outcome, which was significant using the MR-Egger method ($b = 1.27$, $se = 0.48$, $p = .02$; Supplementary Table 8). Further, the sensitivity analyses using binary estradiol levels for the combined pre- and postmenopausal, premenopausal-only, and postmenopausal-only females from the UKB were not significant for any of the outcomes (Supplementary Table 9).”

And the results of the multivariable Mendelian randomization analyses using the new GWAS summary statistics are included (page 11) as follows:

“In multivariable analyses, genetically-predicted BMI and continuous estradiol levels in the combined pre- and postmenopausal sample were not significantly associated with brain age gap, Alzheimer’s disease, or depression as outcomes (Supplementary Table 12). Results were consistent when using binary estradiol levels as an exposure.”

Further, we have rewritten our discussion (page 12-13) to include the results of the newly run GWAS as follows:

“Further, we ran and annotated a GWAS on estradiol levels in females from the UKB, to avoid sample overlap. We identified two significant genomic loci, of which one (on chromosome 17) has been previously reported by a study on estradiol levels in females in the UKB²⁴. This locus is near the sex-hormone binding globulin (SHBG) gene, which is closely linked to concentrations of SHBG, a glycoprotein that binds to estradiol and testosterone^{23,24,32}. Additionally, the present study identified a significant locus on chromosome 19 that has been associated with white blood cell count and height, but not estradiol levels (see NHGRI-EBI GWAS Catalog³²). Differences between the present study and previous studies may partly be traced back to differences in covariates and sample selection. The GWAS of the present study

might be limited by the included covariates. For instance, while we controlled for menopausal status as well as history of oral contraceptive use, HRT use, hysterectomy, and oophorectomy, menstrual cycle phase was not considered due to the inclusion of pre- and postmenopausal females, potentially influencing the findings.”

4. It would be useful to the reader if the paper outlined more specifically how proxies were identified, particularly (if it were possible) with this added to the associated code files. The paper describes in the supplementary material that the proxy variants were searched for using the LDProxy Tool from LDlink, but a more detailed explanation of how this identifies proxies would benefit readers, especially given the large discrepancies in SNPs identified in the depression and alzheimer's datasets when the same exposures are being used.

Response: We thank the reviewer for this comment and acknowledge that more details were needed on how proxy variants were identified. We have now elaborated on the criteria we used for identifying proxy variants in the revised supplements (Supplemental Note 2, page 5) as follows:

“Proxy variants were searched for using the LDProxy Tool from LDlink. The respective instrumental variable was entered in the search using the Great British reference population (GBR). When available, resulting proxy variants with a minimum linkage disequilibrium (LD) R^2 of .60 were ordered according to their R^2 and the proxy variant with the highest R^2 was used.”

5. The paper is very long, with an especially long discussion with subheadings.

Response: We thank the reviewer for this comment and have shortened the discussion where possible and we have removed the subheadings from the discussion in the revised manuscript. For instance, we have shortened the section discussing the GWAS of brain age gap in females (page 12) to one paragraph. We have also shortened the section on GWAS of estradiol levels in females (page 12-13), with the new focus on the continuous estradiol phenotype. Further, in line with minor comment 3, we have reduced the descriptions of the findings, including in the first paragraph discussing the Mendelian randomization analyses (page 13) as well as in the last paragraph of the discussion (page 16). However, some points in the discussion required more details and needed to be extended.

Minor comments

1. Figure 1a(3) is a low quality table which cannot be read when zoomed in.

Response: We have now included the figure in higher quality (page 26).

2. Table 1, by crossing two pages and including a large amount of information is difficult to read and may be more effective in the supplement with landscape formatting. It would also be useful (given the differences in numbers of SNPs used) to show what the source of the data for the different instruments was by rows.

Response: Following this suggestion, we have changed the table to landscape format in the revised manuscript (page 20-21) and we have moved the variables used for sensitivity and supplemental analyses to the revised supplements (Supplementary Table 1; page 28-29), to shorten the table in the main manuscript. Further, we have changed the column name showing the source of the data to *Data Source* to avoid confusion.

3. The paper could attempt to explain their findings further in the discussion. There is lots of description of the findings. For example, the attenuation of the significant causal association between age at menarche and depression once BMI was included is described, but an explanation as to why this might be would benefit readers seeking to make biological sense of Mendelian randomisation analyses.

Response: In the revised manuscript, we have reduced the descriptions of the findings to avoid repetitions. To this end, we have shortened the descriptions of the Mendelian randomization results in the first paragraph discussing these analyses (page 13) as well as in the last paragraph of the discussion (page 16).

We have attempted to further explain our findings in the revised manuscript. For instance, we have elaborated on the attenuation of the significant causal association between age at menarche and depression once BMI in the discussion of the revised manuscript (page 14-15) as follows:

“Previous Mendelian randomization studies have reported a causal link between a younger age at menarche and an increased risk for depression in adolescents and adults^{16,17,19,21,22}. In the multivariable Mendelian randomization analyses of the present study, the attenuation of the significant effect when including BMI suggests potential pleiotropy, whereby the genetic

variants included in the analysis with age at menarche influence depression through BMI. BMI is known to be linked to an earlier onset of puberty and, in line with the present study, has previously been highlighted as a relevant confounder in the causal relationship between a younger age at menarche and depression^{16,19}. Furthermore, previous studies have discussed possible age-dependent effects of age at menarche on depression^{16,19}, warranting exploration to explain inconsistent findings.”

And we have further explained our null findings (page 13-14) as follows:

“However, it must be mentioned that some of the additionally analyzed exposure variables, such as HRT use and oral contraceptive use, might not have a substantial genetic component, potentially leading to null findings in the present study. Furthermore, non-linear and age-dependent effects, as have been suggested for associations of number of childbirths and HRT use with brain health^{7,36}, should be investigated. For instance, previous studies highlight the importance of timing, duration, and formulation of HRT use³⁶. These aspects were not captured by the variable used for HRT use in the present study, possibly resulting in null findings. Further, the GWAS of history of hysterectomy did not take important influencing factors into account, such as the medical reason for the procedure, the timing of procedure (i.e., before or after menopause), and whether an oophorectomy also took place. These limitations can be traced back to the use of GWAS from consortia, which have advanced research through publicly available, large-scale datasets, however, are often designed in a standardized matter across sexes and may not consider potentially relevant covariates for specific variables or samples. Future research is needed to disentangle genetic and environmental contributions to these variables and further assess their causal associations with brain and mental health.”

4. It would seem prudent to discuss fully, when covering the limitations, that for many of the MR analyses, the small number of SNPs means that the power of these analyses would be quite low.

Response: In line with this comment, we have included this limitation in the discussion of the revised manuscript (page 14) as follows:

“Further, power issues might be present in the analyses with few instrumental variables, especially in the sensitivity analyses using methods with lower power, such as MR-Egger, and in the presence of weak instrument bias³⁸.”

Reviewer #2 (Remarks on code availability):

Code appears to be standard for this type of analysis, but it would be useful to have the code that was used to identify any proxies used in this analysis.

Response: As we did not use code to identify proxy variants, we decided against adding this explanation to the code. However, we have added a more detailed explanation of how proxy variants were identified in the revised supplements (Supplemental Note 2, page 5) as follows:

“Proxy variants were searched for using the LDProxy Tool from LDlink. The respective instrumental variable was entered in the search using the Great British reference population (GBR). When available, resulting proxy variants with a minimum linkage disequilibrium (LD) R^2 of .60 were ordered according to their R^2 and the proxy variant with the highest R^2 was used.”

Reviewer #3 (Remarks to the Author):

This was an interesting and well written study by Oppenheimer et al looking at estradiol exposure and the female brain and mental health. The main result is that estradiol exposure is not causing depression or altering the female brain age gap as might be expected based on the observational literature. The methodology used is generally sound although I do have a few questions:

General response: We thank the reviewer for taking the time to evaluate our manuscript, and for the helpful suggestions for improvements. Please find our responses to each comment below, and the corresponding edits highlighted in yellow in the uploaded copy of the revised manuscript file.

1. Is UKB the best option for a GWAS of estradiol levels? Given the age range of the group many women have gone through menopause and this means that most women in the study have a reading outside the detectable limit. Does the binary metric you have used effectively capture menopause rather than hormone levels? This was previously discussed in Ruth et al: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7025895/> I'd certainly question how accurate the exposure was at solely capturing estradiol levels. For comparison could you look at estradiol levels in men, which did seem to somewhat heritable in men?

Response: We thank the reviewer for this comment and while we agree that the age range of the UKB is not ideal for estradiol level measurements in females, it is the largest dataset available to date. Following this comment as well as comment 1 of reviewer #1 and comment 1 of reviewer #2, we have now replaced the binary estradiol phenotype with the continuous estradiol phenotype in the main manuscript and have moved the analyses with the binary estradiol phenotype into the supplements. Further, we have replicated the results on estradiol levels in additional samples. Specifically, we now include GWAS summary statistics conducted in a sample of postmenopausal females from the LIFE-Adult and LIFE-Heart study (Pott et al., 2019) as well as samples of only premenopausal females and only postmenopausal females from the UKB, using both the continuous and binary phenotypes (Haas et al., 2022). Additionally, following the suggestion of comparing estradiol levels in males, we have replicated the analyses using estradiol levels in a male-only sample from the UK Biobank

(<http://www.nealelab.is/uk-biobank>). To this end, we have obtained the GWAS summary statistics of the outcome variables (depression and Alzheimer's disease) in male-only samples and we have run and annotated a male-specific GWAS on brain age gap. The results were consistently non-significant across samples. By including multiple additional samples, we are confident that we substantially increased the robustness of our results.

We have now included these additional analyses in the methods section of the revised manuscript under *Samples and Data Sources: Available Summary Statistics* (page 17) as follows:

“To ensure robustness to effects of menopausal status previously shown to impact GWAS results²⁴, several datasets were included for estradiol levels. The following samples from the UKB were included: a combined sample of pre- and postmenopausal females, a sample of only premenopausal females, and a sample of only postmenopausal females²⁴. Main analyses in each sample were performed using GWAS of continuous estradiol levels. Sensitivity analyses were conducted in each sample using a binary approach examining estradiol levels above and below the detection limit²⁴ to increase sample sizes. Further, analyses on continuous estradiol levels were replicated in an independent sample of postmenopausal females from the LIFE-Adult & LIFE-Heart studies⁴⁵ and a male-only sample from the UKB⁴⁶.”

Further, we have included these datasets and analyses in Table 2 (page 20-21) and Supplementary Table 1 (Supplementary Information page 28-29). The results of the analyses are mentioned in the revised manuscript under *Mendelian Randomization Analyses* (page 7-8) as follows:

“No significant causal relationships were found between estradiol levels, using the continuous measures as the exposure, and brain age gap, Alzheimer's disease, and depression as outcomes (Table 1). The results were consistently non-significant across the combined, premenopausal-only, and postmenopausal-only samples from the UKB. Similarly, the replication analyses in the independent postmenopausal sample from the LIFE studies were non-significant (Table 1). All results remained non-significant across the estimation methods, except for the analysis with estradiol levels in the combined pre- and postmenopausal sample as an exposure and Alzheimer's disease as an outcome, which was significant using the MR-Egger method ($b = 1.27$, $se = 0.48$, $p = .02$; Supplementary Table 8). Further, the sensitivity analyses using binary estradiol levels for the combined pre- and postmenopausal, premenopausal-only, and postmenopausal-only females from the UKB were not significant for any of the outcomes

(Supplementary Table 9). For the male samples, no significant associations were found between estradiol levels and brain age gap, Alzheimer's disease, and depression (Supplementary Table 9). The results remained non-significant across the estimation methods (Supplementary Table 8)."

We hope to see an increase in large-scale datasets with reliable estradiol measurements in wider age ranges in the future. We mention this in the discussion (page 13) of the revised manuscript as follows:

"Many large-scale databases, including the UKB, are conducted on aging cohorts and are confounded by survivor and healthy volunteer biases³⁴. Therefore, we conducted sensitivity analyses across different samples and approaches. Nevertheless, the present study is limited by the available data on estradiol levels and further highlights the need for large-scale, precise measurements conducted in diverse samples and age groups under consideration of female-specific variables – a data gap which has been repeatedly identified³⁵."

2. When you only have 8 or 9 exposure loci how reliable is MR Egger? I believe that when 10 or less instruments it is a fairly imprecise tool.

Response: We thank the reviewer for this comment. In general, MR-Egger has a lower power than the inverse-variance weighted (IVW) method (Bowden et al., 2015). The power increases with a larger number of instrumental variables, however, to the best of our knowledge, there are no set guidelines on the minimum number of required instruments (Bowden et al., 2015). Further, power issues may especially constitute a problem when there is weak instrument bias (generally defined as $F < 10$). Weak instrument bias was only present in the analyses including number of childbirths as an exposure with depression and recurrent depression as outcomes as well as the multivariable analyses, but not for any of the other analyses.

We have now mentioned the possible power issues in the discussion (page 14) of the revised manuscript as follows:

"Further, power issues might be present in the analyses with few instrumental variables, especially in the sensitivity analyses using methods with lower power, such as MR-Egger, and in the presence of weak instrument bias³⁸."

In general, MR-Egger is a useful method for sensitivity analyses. When estimates between the IVW and MR-Egger methods differ substantially, caution in the interpretation of effects is warranted (Bowden et al., 2015). In the present study, the MR-Egger method agreed with the IVW method throughout analyses, with the exception of the analysis using age at menopause as an exposure and depression as an outcome and the analysis using age at menarche as an exposure and depression and recurrent depression as outcomes. These analyses included 46, 149, and 136 instrumental variables, respectively, and did not suffer from weak instrument bias ($F = 82.9$; $F = 69.4$; $F = 70.4$, respectively).

3. What do the betas in table 2 reflect for estradiol? Is it based on the binary metric? Therefore should convert to a doubling in genetic liability for a binary exposure? <https://pubmed.ncbi.nlm.nih.gov/30039250/>

Response: We thank the reviewer for this comment and apologize for the confusion. We log-transformed the odds ratios of the binary variables before conducting the Mendelian randomization analyses. This was done in order to obtain comparable beta-values across results. Thus, the estimates obtained represent the change in the outcome per unit change in the exposure on the log odds ratio scale. As we report null findings throughout the analyses, we have not attempted to interpret the reported estimates. However, following this suggestion and the mentioned paper by Burgess and Labrecque (2018), we have included the conversion to a doubling in genetic liability (by multiplying the causal estimate by 0.693) for the binary estradiol exposures.

We mention this procedure and the results of the binary exposure variables in the revised supplements in Supplementary Table 9 (page 45) as follows:

*“*Conversion to a doubling in genetic liability for binary exposure variables²¹.”*

4. Given low numbers of exposure in some analysis and the issues with overlap which might limit sample size worth considering using MRlap as a method?

Response: We thank the reviewer for this comment. Following this suggestion, we have included MRlap as a sensitivity analysis for our analyses with sample overlap.

We mention MRlap in the methods section of the revised manuscript under *Two-Sample Mendelian Randomization Analyses: Univariable Analyses* (page 25) as follows:

“Using the MRlap package (v0.0.3.2)⁶⁹, MRlap was performed for analyses with sample overlap. MRlap corrects for potential biases arising from sample overlap, weak instruments, and winner’s curse (i.e., the overestimation of effects in discovery GWAS)⁶⁹.”

Further, in the results section of the revised manuscript under *Mendelian Randomization Analyses* (page 8), we have mentioned the MRlap results as follows:

“Further, when using MRlap to correct for potential biases resulting from sample overlap, the results of the analyses with overlapping exposure and outcome samples (where both samples were derived from the UKB, i.e., estradiol levels in premenopausal females, estradiol levels in postmenopausal females, estradiol levels in males, number of childbirths, HRT use, oral contraceptive use, history of hysterectomy, and history of oophorectomy as exposures with brain age gap as an outcome) remained robust, with no significant differences between the IVW estimates and the corrected estimates (Supplementary Table 8).”

The corrected MRlap estimates and p-values for the difference between the observed and corrected effects are reported in the revised supplements in Supplementary Table 8 (pages 37-44). However, we were unable to perform MRlap for continuous estradiol levels measured in the postmenopausal sample, as a negative heritability was estimated for this exposure, likely related to the small sample size ($N = 3,759$). We have mentioned this in the note of the table (page 44) as follows:

*“MRlap²⁰ was only used for analyses with sample overlap. For MRlap, the corrected b , SE , and p are reported²⁰. p_{diff} corresponds to the test for the difference between the observed and the corrected effect²⁰. *MRlap could not be performed for estradiol levels (postmenopausal, continuous) as an exposure due to the negative heritability estimated for the phenotype ($h^2 = -0.0017$ (0.1004)), likely related to the small sample size.”*

5. Did you look at the potential pleiotropic nature of the variants included as exposure? For example, did you look them up in the GWAS catalogue.

Response: We computed Cochran’s Q as a measure of heterogeneity, which indicates potential presence of pleiotropy when reaching significance (Burgess et al., 2017). We further used the MR-Egger, weighted median, simple mode, and weighted mode estimation methods as

robustness checks for all analyses, which make varying assumptions regarding pleiotropy (Burgess & Thompson, 2021). As we report null findings throughout the analyses, we did not attempt to investigate potential pleiotropy in more detail.

However, following this comment we have decided to more closely examine potential pleiotropy in the GWAS we conducted on estradiol levels (continuous and binary) as exposure variables. We have now included Supplementary Tables 14 and 15 (page 54-65) listing the traits related to the independent significant SNPs and the mapped genes found, as reported in the GWAS catalog.

We mention this in the methods section of the revised manuscript under *GWAS Procedure: Post-GWAS Annotations* (page 23-24) as follows:

“For estradiol levels, independent significant SNPs were linked to traits as reported in the GWAS Catalog³² to facilitate investigation of potential pleiotropy of these variables used as exposures in the Mendelian randomization analyses.”

Reviewer #3 (Remarks on code availability):

Code is clear and easy to follow. Well annotated and from a MR perspective looks sensible.

Response: We thank the reviewer for this comment on the code.

References

- Bowden, J., Davey Smith, G. & Burgess, S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *International Journal of Epidemiology* **44**, 512–525 (2015).
- Burgess, S., Bowden, J., Fall, T., Ingelsson, E. & Thompson, S. G. Sensitivity Analyses for Robust Causal Inference from Mendelian Randomization Analyses with Multiple Genetic Variants. *Epidemiology* **28**, 30–42 (2017).
- Burgess, S. & Labrecque, J. A. Mendelian randomization with a binary exposure variable: interpretation and presentation of causal estimates. *Eur J Epidemiol* **33**, 947–952 (2018).
- Day, F. R. *et al.* Large-scale genomic analyses link reproductive aging to hypothalamic signaling, breast cancer susceptibility and BRCA1-mediated DNA repair. *Nat Genet* **47**, 1294–1303 (2015).
- Elsworth, B. *MRC-IEU Consortium*. <https://gwas.mrcieu.ac.uk>.
- Faubion, S. S. *et al.* Long-term health consequences of premature or early menopause and considerations for management. *Climacteric* **18**, 483–491 (2015).
- Gilsanz, P. *et al.* Reproductive period and risk of dementia in a diverse cohort of health care members. *Neurology* **92**, (2019).
- Gong, J., Harris, K., Peters, S. A. E. & Woodward, M. Reproductive factors and the risk of incident dementia: A cohort study of UK Biobank participants. *PLoS Med* **19**, e1003955 (2022).
- Haas, C. B., Hsu, L., Lampe, J. W., Wernli, K. J. & Lindström, S. Cross-ancestry Genome-wide Association Studies of Sex Hormone Concentrations in Pre- and Postmenopausal Women. *Endocrinology* **163**, bqac020 (2022).
- Huang Y., Wu M., Wu C., *et al.* Effect of hysterectomy on ovarian function: a systematic review and meta-analysis. *J Ovarian Res* **16**, 35 (2023).
- Loh, P.-R., Kichaev, G., Gazal, S., Schoech, A. P. & Price, A. L. Mixed-model association for biobank-scale datasets. *Nat Genet* **50**, 906–908 (2018).
- Moorman P.G., Myers E.R., Schildkraut J.M., Iversen E.S., Wang F., Warren N. Effect of hysterectomy with ovarian preservation on ovarian function. *Obstet Gynecol* **118**, 1271–1279. (2011)
- Pott, J. *et al.* Genetic Association Study of Eight Steroid Hormones and Implications for Sexual Dimorphism of Coronary Artery Disease. *The Journal of Clinical Endocrinology & Metabolism* **104**, 5008–5023 (2019).
- Schindler, L. S. *et al.* Associations between abdominal adipose tissue, reproductive span, and brain characteristics in post-menopausal women. *NeuroImage: Clinical* **36**, 103239 (2022).

Reviewer #1 (Remarks to the Author):

In this revised manuscript the authors have been very responsive to this reviewer's prior comments. While I still have questions regarding whether the MR approach is the most appropriate technique for examining the questions addressed in this study, the manuscript is now appropriate for publication in my opinion. I hope that this study will stimulate a healthy debate within the field.

General response: We thank the reviewer for taking the time to re-evaluate our manuscript.

Reviewer #2 (Remarks to the Author):

It seems that there have been some substantial changes to the underlying instrument used in this study, and I think that that puts in on a much more secure footing that avoids the previous issues of confounding.

General response: We thank the reviewer for taking the time to re-evaluate our manuscript, and for the helpful suggestions for further improvements. Please find our responses to each comment below, and the corresponding edits highlighted in yellow in the uploaded copy of the revised manuscript file.

1. However, now that the main analysis is the 2 loci from the continuous exposure, I wonder to what extent this can be described as "life-time exposure", given that most of the sample is post-menopausal.

Response: We thank the reviewer for this comment. We chose the term “lifetime exposure” to emphasize that the method used captures constant effects based on genetic variants randomly allocated at conception, rather than time-varying effects and fluctuations. Further, we replicated our findings in a premenopausal sample and a male sample, to test the robustness outside of the sample that mostly consisted of postmenopausal females. However, we now highlight this limitation in the revised manuscript in the discussion section (pages 18-19) as follows:

“Due to the methods used, the present study was unable to examine potential effects of estradiol fluctuations or susceptible periods to estradiol exposure throughout the female lifespan. Therefore, we used the phrase lifetime estradiol exposure. However, the sample used for our main analysis on estradiol levels consisted predominantly of postmenopausal females, thus, likely not capturing effects across the full lifespan. Ideally, to capture time-varying effects, multivariable Mendelian randomization should be applied using GWAS summary statistics across multiple time points⁵¹. For instance, estradiol levels should be assessed repeatedly across the female lifespan.”

2. The other effect is that now the results section is quite confusing, with a number of different scores being proposed and reported in table 1 - there are now three possible samples (pre, post and combined) and two possible measures (continuous and binary) and that's just in UK Biobank, with the LIFE samples as a addition. This made it hard to track the numbers of signals through the paper - especially as the authors have reported some loci as genome-wide significant, mentioned multiple signals at these loci, but then used in some cases a lower threshold to determine instrument selection. It would be useful to have a table somewhere where just the GWAS number of GWAS signals were reported, and separately the number of instruments that were used. I'd also be tempted to suggest splitting Table 1 into one table with just the Estradiol results. And the other phenotypes separately, as more detail about the Estradiol instruments could then be given. Also, while a lower threshold is mentioned in the methods section, I think that this could have been briefly mentioned in the results section to make the article easier to follow.

Response: We thank the reviewer for this comment. To make the thresholds and instrumental variables used for each analysis more clear, we have now included a new table, Supplementary Table 4 (revised supplements pages 36-39), which offers an overview of the threshold used, number of instrumental variables, and instrument strength for each analysis.

Further, as suggested, we have now split Table 1 into two tables in the revised manuscript. Table 1 (pages 9-10) displays the results of the univariable Mendelian randomization analyses using continuous estradiol levels as an exposure across the different samples. Table 2 (pages 12-13) displays the univariable Mendelian randomization analyses using the factors related to lifetime estradiol exposure as exposures. Further, in both tables, we have now included a column for the thresholds used to select instrumental variables for each respective analysis. Additionally, we mention the lower threshold used for the selection of instrumental variables for the continuous estradiol levels in the results section of the revised manuscript (page 7) as follows:

“Across the samples used for continuous estradiol levels, few instrumental variables were identified at the genome-wide significance threshold, therefore, a lower threshold ($p < 5 \times 10^{-6}$) was applied for selecting instrumental variables (see Supplementary Note 3 and Supplementary Table 4 for details).”

3. I thank the authors for the clarification of the method to find proxies, but it still seems a bit surprising that the SNP numbers are so low for Depression compared to the other phenotypes.

Response: We thank the reviewer for this important observation. The low number of SNPs for depression compared to the other phenotypes resulted from the low number of variants available in the summary statistics of the depression GWAS (~ 6.5 million SNPs) compared to the brain age gap GWAS (~ 12.2 million SNPs) and the Alzheimer's disease GWAS (~ 38.0 million SNPs). This was the largest GWAS of diagnosed depression available to us (while avoiding sample overlap) when we first ran the analyses, however, we have recently conducted a larger, female-only GWAS meta-analysis on diagnosed major depressive disorder (https://doi.org/10.31219/osf.io/tjfd9_v1). Therefore, we have rerun our Mendelian randomization analyses using this new GWAS meta-analysis, which more than triples the number of variants for our instrument search (~ 19.9 million SNPs) and substantially increases the sample size to 329,476 females, providing much greater statistical power for our analyses. The use of the new GWAS meta-analysis, which includes the GWAS we originally used (Blokland et al., 2023) as well as data from the UK Biobank (UKB; Sudlow et al., 2015) and the Norwegian Mother and Child Cohort Study (MoBa; Magnus et al., 2016), is described in the methods section of the revised manuscript in Table 3 (page 23). The results largely replicated our previous findings and are described in the results section of the revised manuscript (page 7) as follows:

“No significant causal relationships were found between estradiol levels, using the continuous measures as the exposure, and brain age gap, Alzheimer's disease, and depression as outcomes (Table 1). The results were consistently non-significant across the combined, premenopausal-only, and postmenopausal-only samples from the UKB. Similarly, the replication analyses in the independent postmenopausal sample from the LIFE studies were non-significant (Table 1).”

As well as in the results section of the revised manuscript on page 11 as follows:

“No significant associations were found between the exposures reproductive span, age at menopause, and number of childbirths and the outcomes brain age gap, Alzheimer's disease, and depression (Table 2). Consistently, no significant effects were found in the supplementary analyses using oral contraceptive use, HRT use, history of hysterectomy, and history of

oophorectomy as exposure variables (Supplementary Note 7 and Supplementary Table 8). All results remained robust across the estimation methods. Similarly, age at natural (non-surgical) menopause as an exposure was not significant in the sensitivity analysis for Alzheimer's disease as an outcome (Supplementary Table 11). A significant association was found for age at menarche with depression as an outcome ($b = -0.09$, $se = 0.04$, $p = .04$), however, this result did not remain significant after adjusting for multiple comparisons ($p_{FDR} = .76$) and was not significant when using any of the other estimation methods (Figure 3)."

The results for the multivariable Mendelian randomization analyses are described in the results of the revised manuscript (page 14) as follows:

"For depression as an outcome, continuous estradiol levels in the combined pre- and postmenopausal sample were not significant, but BMI was significant ($b = 0.14$, $se = 0.05$, $p = .01$, $p_{FDR} = .30$) in multivariable analyses. This was not robust when using multivariable MR-Egger; however, the result was consistent when using binary estradiol levels as an exposure (BMI: $b = 0.15$, $se = 0.05$, $p = .003$, $p_{FDR} = .27$). Furthermore, when including BMI, the association between age at menarche and depression as well as the association between age at menarche and recurrent depression were not significant. This remained consistent when using multivariable MR-Egger. No significant heterogeneity was found in any of the analyses, except for the analysis using age at menarche and BMI as exposures and depression as an outcome (Supplementary Table 14)."

Further, we updated Table 1 (pages 9-10), Table 2 (page 12-13), and Figure 3 (page 11) in the revised manuscript to include the new depression GWAS. The same was done for Supplementary Table 2 (page 34), Supplementary Table 8 (pages 44-47), Supplementary Table 9 (pages 48-52), Supplementary Table 10 (pages 53-56), Supplementary Table 13 (pages 63-66), and Supplementary Table 14 (pages 67-69).

To ensure the validity of our findings from this new GWAS, we used multiple approaches to address potential sample overlap. Crucially, our primary findings remained consistent across the main analysis using the GWAS meta-analysis, sensitivity analyses on a subsample without sample overlap (Blokland et al., 2023), and when using MRlap for analyses with sample overlap. The sensitivity analyses are described in the introduction of the revised manuscript (page 5) as follows:

“As further sensitivity analyses, we use risk for recurrent depression as an outcome, to assess potential influences of disease burden, as well as a subsample of the depression GWAS to avoid sample overlap.”

The sample overlap is described in the methods section of the revised manuscript (page 22) as follows:

“Sample overlap was avoided for all analyses using Alzheimer’s disease and recurrent depression as outcomes, as well as the sensitivity analyses in the depression⁵⁴ subsample excluding UKB. Sample overlap with brain age gap and depression as outcomes was avoided for age at menarche and age at menopause as exposures by using datasets that did not include the UKB sample. Further, sample overlap with brain age gap as an outcome was avoided for estradiol levels (combined pre- and postmenopausal samples) and reproductive span as exposures by excluding the MRI sample from the UKB. However, sample overlap could not be avoided for the analyses using continuous estradiol levels in the combined pre- and postmenopausal sample and reproductive span as exposures and depression as an outcome (maximum sample overlap: 10.53% for estradiol levels and 36.91% for reproductive span). Further, there was sample overlap for continuous estradiol levels in the premenopausal sample and in the postmenopausal sample with brain age gap and depression (maximum sample overlap: 43.25% for premenopausal females with brain age gap as an outcome; 10.03% for premenopausal females with depression as an outcome; 26.31% for postmenopausal females with brain age gap as an outcome; 1.14% for postmenopausal females with depression as an outcome). For number of childbirths, there was a maximum sample overlap of 5.70% with brain age gap as an outcome and 82.19% with depression as an outcome. Further, sample overlap could not be avoided for the supplementary analyses using factors related to exogenous hormone use and health-related procedures and the sensitivity analyses using estradiol levels in males with brain age gap and depression as outcomes, as well as using binary estradiol levels with depression as an outcome (see Supplementary Table 2).”

And in the methods section of the revised manuscript (pages 20-21) as follows:

“Further, as some of the analyses using depression²⁷ as an outcome had sample overlap, these analyses were replicated using a subsample of the depression GWAS that did not include the UKB sample⁵⁴.”

The results of the sensitivity analyses are described in the revised supplements in Supplementary Table 12 (pages 58-62) and in the results section of the revised manuscript (page 14) as follows:

“Further, the sensitivity analyses using the depression subsample excluding UKB for analyses with sample overlap replicated the findings of the main analyses, with no significant associations across analyses (Supplementary Table 12).”

Additionally, we ran a male-only GWAS meta-analysis for diagnosed depression using the same approach as the female-only GWAS meta-analysis ($N = 262,747$) and reran the sensitivity analysis for estradiol levels in males. This GWAS is described in the introduction of the revised manuscript (page 5) as follows:

“To this end, we run a male-specific GWAS on brain age gap using data from the UKB and a GWAS meta-analysis including data from the UKB, the Norwegian Mother and Child Cohort Study (MoBa; a population-based pregnancy cohort study conducted by the Norwegian Institute of Public Health; see Magnus et al. (2016)²⁵ for details), and a previously conducted GWAS.”

As well as in Supplementary Table 1 (pages 32-33) and in the methods section of the revised manuscript (page 21) as follows:

“For depression in the male-only sample, a GWAS meta-analysis was run in line with the female-only GWAS²⁷ including data from the UKB¹⁰, MoBa²⁵, and a previously conducted GWAS received from the authors upon request⁵⁴ (see Supplementary Notes 1, 2, and 5 for details).”

Details are given in the revised supplements in Supplementary Note 1 (page 5) as follows:

“Exclusion of Participants from Male-Only Depression GWAS Meta-Analysis

In line with the female-only depression GWAS meta-analysis¹, GWAS were run on males with White European ancestry from the UK Biobank (UKB)² and the Norwegian Mother and Child Cohort Study (MoBa)³. The UKB is a prospective population-based study from the United

Kingdom, encompassing over 500,000 participants aged 40-69 that were recruited between 2006 and 2010³. This research has been conducted using the UKB Resource under Application Number 27412. MoBa is a population-based pregnancy cohort study conducted by the Norwegian Institute of Public Health³. Participants were recruited from all over Norway from 1999-2008³. The women consented to participation in 41% of the pregnancies³. The cohort includes approximately 114,500 children, 95,200 mothers, and 75,200 fathers³. MoBa is regulated by the Norwegian Health Registry Act³. The present study was approved by the Regional Committees for Medical and Health Research Ethics (2016/1226/REK).

Inclusion and exclusion criteria were chosen following previous studies^{1,4}. Cases were defined as having a lifetime (primary or secondary for UKB; following the Norwegian Patient Registry for MoBa) diagnosis of a depressive episode or recurrent depressive episode, according to the ICD-10 (F32 or F33). Cases with a lifetime (primary or secondary) diagnosis of schizophrenia, schizotypal, or delusional disorder (F20-F29), mania or bipolar disorder (F30 or F31), or personality disorder (F40 or F61) were excluded. Further, controls with a lifetime (primary or secondary) diagnosis of any mood disorder (F30-F39), schizophrenia, schizotypal, or delusional disorder (F20-F29), or personality disorder (F40 or F61) were excluded. Finally, $n = 9,413$ cases and $n = 176,243$ controls were included from the UKB and $n = 2,470$ cases and $n = 45,628$ controls were included from MoBa. From the PGC and iPSYCH samples ($n = 10,194$ cases and $n = 18,799$ controls), GWAS summary statistics of diagnosed depression in males were received from the authors upon request⁵. The following linkage disequilibrium score (LDSC; v2.0.0) heritability^{6,7} estimated were computed: UKB $h^2 = 0.02$ (SE = 0.00), MoBa $h^2 = 0.11$ (SE = 0.05), and PGC and iPSYCH $h^2 = 0.07$ (SE = 0.02). Using LDSC genetic correlations, the UKB and MoBa GWAS correlated at $r_g = 0.87$ (SE = 0.24, $p = 2.00e^{-4}$), the UKB and PGC and iPSYCH GWAS correlated at $r_g = 1.05$ (SE = 0.21, $p = 7.86e^{-7}$), and the MoBa and PGC and iPSYCh GWAS correlated at $r_g = 0.67$ (SE = 0.31, $p = .03$)."

And in Supplementary Note 2 (page 6) as follows:

"Note 2. Procedure of Male-Only Depression GWAS Meta-Analysis.

The UKB v3 imputed genetic data was used for the UKB GWAS, which has been genotyped, extensively quality controlled, and imputed by the UKB genetics team⁸. The imputed genetic data from the MoBaPsychGen pipeline v.1⁹ was used for the MoBa GWAS. Both GWAS were performed using REGENIE v4.1¹⁰. For the association analysis, we retained only autosomal variants with a minor allele count > 20 and an imputation information score > 0.80.

Covariates included age, the first twenty genetic principal components and genotyping batch (only for MoBa). METAL (version 2020-05-05; <https://github.com/statgen/METAL>)¹¹ was used for the inverse variance-based meta-analysis of the GWAS summary statistics from the UKB, MoBa, and PGC and iPSYCH samples.”

The results of the GWAS meta-analysis are described in the revised supplements in Supplementary Note 5 (page 12) as follows:

“Note 5. Results for Male-Only Depression GWAS Meta-Analysis.

The male-only depression GWAS meta-analysis identified no genome-wide significant SNPs (see Supplementary Figures 3 and 8). The most significant SNP was located on chromosome 11 (rs145678014; $p = 0.81 \times 10^{-8}$).”

The results replicated the previous findings and are described in the revised supplements in Supplementary Table 10 (pages 53-56). The results of the sensitivity analysis using the male-only depression GWAS by Blokland et al. (2023), avoiding sample overlap, are described in the revised supplements in Supplementary Table 12 (pages 48-62).

4. With regard to the limitations of UKB, a major issue is the overall rate of detection of the hormone, while this is mentioned on line 348, it would be useful if the authors could discuss how they think it would have specifically impacted their study.

Response: We thank the reviewer for this comment and agree that this is an important issue to highlight. Therefore, we have now elaborated further on this in the discussion section of the revised manuscript (page 16) as follows:

“Issues concerning estradiol measurements in the UKB have been previously discussed, including a potential bias towards the detection of loci associated with menopause due to the age of the participants and the substantial number of measurements below the detection limit^{23,24,41}. This may influence the findings of the present study, as the detected loci of the GWAS as well as the selected instruments for estradiol levels may be associated with related traits. This raises the possibility of confounding from horizontal pleiotropy¹⁵, where the genetic instruments influence the outcome through pathways other than the exposure (estradiol), such as via menopausal status or SHBG levels.”

Further, in the discussion section (page 16), we mention the following:

“Nevertheless, the present study is limited by the available data on estradiol levels and further highlights the need for large-scale, precise measurements conducted in diverse samples and age groups under consideration of female-specific variables – a data gap which has been repeatedly identified⁴³.”

Feedback on the responses to reviewer 3's previous comments:

The authors responses to points 1-3 are all comprehensive and sufficient.

General response: We thank the reviewer for taking the time to re-evaluate our manuscript and reviewer 3's comments, and for the helpful suggestions for further improvements. Please find our responses to each comment below, and the corresponding edits highlighted in yellow in the uploaded copy of the revised manuscript file.

1. Point 4. Regarding the use of MRlap to address any possible issues of sample overlap, it appears from Supplementary Table 8 that MRlap was only used for a subset of the analyses. The authors correctly suggest that MRlap is only of benefit where there is the possibility of sample overlap. However, for a number of the traits that are considered in this paper, there is more than one data set used (as described in Table 2). And for some of the phenotypes listed in Supplementary Table 8, there could be sample overlap depending on the data set chosen for the analysis. I'm aware that the authors have made some effort to try to have independent samples where this was possible, but to make this clear to the reader, could a signifier (probably first author and year) be used to indicate the source in Supplementary Table 8. It might also be helpful to do this for age at menarche and age at menopause in Table 1 in the main text.

Response: We thank the reviewer for this comment and agree that further clarification of the datasets used is helpful. Therefore, we have now included additional columns ("*Exposure GWAS and Data Source*" and "*Outcome GWAS and Data Source*") in the revised supplements in Supplementary Table 9 (formerly Supplementary Table 8; pages 48-52) and in the revised manuscript in Table 1 (pages 9-10) and Table 2 (pages 12-13). These columns include the author, year, and sample of the exposure and outcome datasets used for each analysis. For consistency, we further included these columns in the revised supplements in Supplementary Table 8 (pages 44-47), Supplementary Table 10 (pages 53-56), Supplementary Table 11 (page 57), Supplementary Table 12 (pages 58-62), and Supplementary Table 13 (pages 63-66).

2. Point 5. While this data is very comprehensive, the formatting of Table 14 and 15 could be better. It's not clear what is different between the columns "Independent Significant SNP" and "SNP". The caption should explain this more clearly. There are also multiple

rows where information is exactly duplicated in other rows. I'm also not sure that the added sentence on pages 23/24 is quite sufficient. The authors spend some time in the discussion talking about the overlap with SHBG, so it would make sense to also mention it here.

Response: We thank the reviewer for this comment. The duplicated rows in the Supplementary Tables were erroneous and have now been removed. We have further added a description of “Independent Significant SNP” and “SNP” in the captions of Supplementary Tables 15 and 16 (formerly Tables 14 and 15) in the revised supplements (pages 74 and 80), as follows:

“Note. List of single-nucleotide polymorphisms (SNPs) and reported traits from studies included in the GWAS Catalog³⁵ based on independent significant SNPs identified in the GWAS summary statistics of continuous estradiol levels (see Supplementary Table 6). Table created using FUMA¹⁵. SNP: All SNPs from the GWAS Catalog³⁵ in linkage disequilibrium of the identified independent significant SNPs.”

Further, we have elaborated on this in the discussion of the revised manuscript (page 16) as follows:

“The independent significant SNPs identified in our GWAS have been linked to various related traits, including SHBG levels and testosterone levels (see Supplementary Table 15).”

And as follows (page 16):

“Issues concerning estradiol measurements in the UKB have been previously discussed, including a potential bias towards the detection of loci associated with menopause due to the age of the participants and the substantial number of measurements below the detection limit^{23,24,41}. This may influence the findings of the present study, as the detected loci of the GWAS as well as the selected instruments for estradiol levels may be associated with related traits. This raises the possibility of confounding from horizontal pleiotropy¹⁵, where the genetic instruments influence the outcome through pathways other than the exposure (estradiol), such as via menopausal status or SHBG levels.”

Reviewer #2 (Remarks on code availability):

No further comments on the code after last time.

Reviewer #4 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

General response: We thank the reviewer for taking the time to re-evaluate our manuscript.

References

- Blokland, G. A. M. *et al.* Sex-Dependent Shared and Non-Shared Genetic Architecture Across Mood and Psychotic Disorders. *Biological Psychiatry* **91**, 102–117 (2022).
- Magnus, P. *et al.* Cohort Profile Update: The Norwegian Mother and Child Cohort Study (MoBa). *Int. J. Epidemiol.* **45**, 382–388 (2016).
- Sudlow, C. *et al.* UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLoS Med* **12**, e1001779 (2015).