

Brain-cognition relationships and treatment outcome in treatment-resistant late-life depression

Corresponding Author: Dr Aristotle Voineskos

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)

The present manuscript addresses an important research question, examining how neural correlates of executive and memory functions relate to treatment outcomes in older adults with treatment-resistant depression. In a longitudinal study embedded in a multi-center randomized controlled trial, the authors assessed 397 participants at baseline with neuropsychological testing and obtained neuroimaging data (functional connectivity, diffusion metrics, and gray matter structure) in 234 of these. They applied multivariate partial least squares regressions to test associations between cognitive performance and brain measures, finding that 1) better cognition was linked to lower connectivity between default mode and frontoparietal networks as well as within the frontoparietal network; 2) reduced white matter tract integrity across several white matter tracts corresponded to poorer executive performance, and 3) gray matter structure was positively related to cognition. Finally, the authors used regularized elastic net logistic regression to identify baseline biomarkers predictive of remission and observed that poorer cognition and gray matter structure predicted treatment resistance.

I am impressed by the multimodal dataset and the rigor of the analyses: the validation steps (e.g., testing predictive performance of partial least squares or other models in held-out samples), integration of clinically relevant biomarker investigations, and comparison to null models to assess the added predictive value of neuroimaging variables all strengthen the study.

Nonetheless, I have a few points for the authors to address. My major concern pertains to the sample size in predictive analyses: although recruiting older adults with treatment-resistant depression is inherently difficult and the cohort is impressive, the subset used to predict treatment outcomes in step 2 of the treatment protocol is barely 100 participants, which may inflate estimates of model performance (see, for example, <https://doi.org/10.1038/s41386-021-01020-7>). I recommend that the authors report the rates of treatment resistance in Table 1 to provide context for predictive prevalence and explicitly discuss the limited sample size and potential overfitting in the Limitations section. Additionally, it will help readers if the Results clearly explain the different trial steps (e.g., steps 1 and 2) as treatment phases rather than analysis steps; currently this distinction only becomes apparent when inspecting the figure, which may confuse readers initially. There are a few minor issues as well.

- In the introduction, the definitions seem conflated: the first sentences refer to “Treatment-resistant depression (LLD) ...” but it is unclear whether “LLD” was intended to denote late-life depression or treatment-resistant late-life depression.

- For the permutation tests of PLS, please provide exact p-values in the text, especially since many approach the significance threshold (e.g., around 0.03), so readers can judge robustness.

- In Figure 2, the caption describes generalizability as “good,” but the main text characterizes it as “modest,” which seems more accurate given correlations between true and predicted scores are below $r=0.1$. Also, panels 2F and 2H appear to lack regression lines.

- Finally, specify the parcellation atlas used for gray matter analyses (e.g., if FreeSurfer’s Desikan–Killiany atlas plus subcortex), so readers can fully understand the anatomical definitions.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)

In a large cohort of older adults with treatment-resistant late life depression (LLD), the authors examined the neural correlates of cognitive function, and how these associate with prospective treatment outcomes following pharmacotherapy. The study has several notable strengths. This includes a large and robustly-phenotyped sample of older adults with treatment-resistant LLD, along with multi-modal neuroimaging data spanning structural and functional indices, and evaluation of both baseline and clinical trial data. Treatment-resistant late-life depression reflects a unique subset of older adults predisposed to poor medical outcomes and underrepresented in clinical research – which hinders treatment advancement. Studies in this population, asking questions like those posed here, are timely and necessary with potential for broad clinical impacts. However, there are several areas that require attention and further clarification to improve interpretation and impact. Please consider the following comments and suggestions which may help to strengthen the manuscript.

Key Points:

1. Specify which findings were evaluated cross-sectionally so readers can distinguish results derived from baseline vs interventional data.
2. Provide further detail regarding the 'protocolized antidepressant treatment.' If space does not permit a full description of the intervention, stating pharmacotherapy more broadly would aid interpretation so readers can distinguish between other antidepressant treatment modalities (e.g., behavioral interventions).

Abstract:

1. Please report the cognitive status of participants (e.g., unimpaired, MCI).
2. Report the type of intervention participants received.
3. "Education and better structural brain maintenance, but not overall health were associated with better cognition". It's not clear what structural brain maintenance reflects or how this metric is defined. The index for overall health is also not reported. Please specify.

Introduction:

1. Consistency of population and brain network descriptors. Treatment-resistant LLD and LLD are used interchangeably throughout, though as the authors correctly note, treatment resistant LLD is associated with distinct clinical trajectories. They also switch between two descriptions for the same network (frontoparietal and executive control). Please select one for consistency and ease of understanding. Relatedly, the authors use 'LLD' as the acronym for treatment resistant depression in sentence 1 of the introduction.

Methods:

1. Methods state "Exclusion criteria included severe neuropsychiatric conditions such as Parkinson's Disease or schizophrenia." Parkinson's is typically classified as neurodegenerative rather than neuropsychiatric condition. In addition, were those with neurodegenerative conditions also excluded at study start?
2. Relatedly, the cognitive status of participants is difficult to ascertain. The methods state that those with dementia (Short Blessed Test ≥ 10) were excluded. Yet, Figure 1 reports that the sample included subjects with various levels of cognitive function and Table 1 reports that there were 13 subjects with probable dementia included in the sample. Further clarification on the inclusion/exclusion and cognitive status of participants is needed.
3. According to the supplement, there were 212 subs with useable T1 data, but 219 subs with useable DTI data included in the analyses. Often, T1 data is used in the preprocessing of DTI data. It's unclear how this would result in more participants with DTI data included in the analyses than T1, unless T1 wasn't used in the preprocessing. Please clarify.
4. "Among 73 reconstructed tracts, we excluded tracts with more than 3% unusable data and imputed missing data in the remaining 62 tracts." How was imputation performed?

Results:

1. It appears that about 1/3 of the sample underwent cognitive testing and neuroimaging after treatment started (per Supplemental Table 3). To address this, the authors conducted sensitivity analyses excluding participants with neuropsychological testing completed over six weeks after the start of treatment. Can the authors provide justification for selecting 6 weeks after treatment initiation?
2. Related to the above, did the authors run sensitivity analyses focusing only on subjects whose cognitive/neuroimaging data was collected before treatment began? Treatment duration was 10 weeks. If the aim is to examine cognitive/MRI

predictors of treatment response, and the cognitive/MRI predictors are collected after treatment begins, this could significantly impact the findings and what those findings mean. Prior studies have shown changes in resting state connectivity and even modest changes in cognition following pharmacotherapy of brief duration (1-4 weeks). As the authors acknowledge in the discussion, it's possible treatment impacted brain and cognitive markers. Demonstrating that the results persist in the subset of those whose predictors (cognitive performance and brain MRI) were collected before treatment commenced would significantly strengthen the results and interpretation.

3. Given that the study is focusing, in part, on predictors of treatment response, can the authors report in the results the proportion of responders in Step 1 and Step 2?

4. More than 40% of the sample was classified as having MCI. A core aim of the manuscript is to evaluate the neural correlates of cognitive function, and it is very likely that these relationships would differ based on cognitive status, given structural and functional brain changes linked to MCI. It is recommended that the authors examine whether the pattern and directionality of the findings differ based on cognitive status (unimpaired vs. MCI).

5. "Similarly, lower connectivity of different FPN components with each other was also robustly associated with cognitive performance." Specify the directionality (better or worse cognitive performance).

6. Part A of the Trail Making Test is described throughout the manuscript as an executive function test. However, Trails A is widely considered to be a measure of visual scanning and psychomotor/processing speed. Trail Making Test Part B, in contrast, is classified as an executive function measure. The description of Trails A and interpretation of the various results pertaining to this measure should be updated throughout to reflect this.

7. The abstract mentions results pertaining to 'overall health,' but these findings do not seem to be included in the results.

8. "When operationalizing treatment response as change in MADRS scores, a PLS regression with demographic, clinical, cognitive and brain structure data achieved considerable in-sample and hold-out accuracy." Please specify what clinical and demographic data were included. For instance, were sex and age incorporated?

Discussion:

1. The AUC's reported are generally below the thresholds considered clinically useful. This should be acknowledged in the discussion.

2. It appears that white matter (FA) explained much less variance relative to resting state functional connectivity (1.6% WM vs. ~8% with rsFC). Can the authors speculate about why this may be the case? In addition, the Trail Making Test, Part A, was the only cognitive test related to FA. Whereas resting state FC was related to 12 cognitive tests comprising multiple domains. The authors may consider discussing the differences between the white matter and connectivity findings and what this may suggest regarding the neural correlates of cognitive dysfunction in treat-resistant LLD.

3. In Step 1, both cortical thickness and cognition were predictive of remission. In step 2, cortical thickness but not cognitive data predicted remission. Do the authors have theories as to why this might be the case? Could it have something to do with the drug class or characteristics of responders to certain pharmacotherapies but not others?

4. "Cognitive performance and gray matter integrity predicted remission after 10 weeks of bupropion or aripiprazole treatment in the OPTIMUM trial." The authors may consider rephrasing this for accuracy, as a large portion of participants completed cognitive/MRI assessments after drug treatment started (i.e., for many subjects, it was less than 10 weeks of treatment duration after cognitive/MRI data collection).

5. "Our supplementary results support the former hypothesis, given that excluding participants with post-treatment scans and cognitive testing did not substantially change the results." Based on the supplement, it sounds like sensitivity analyses excluded those whose cognitive/MRI data was obtained 6 weeks after treatment began, as opposed to excluding all subjects whose cognitive/MRI data was obtained after treatment started. Please clarify.

(Remarks on code availability)

N/A

Reviewer #3

(Remarks to the Author)

This study investigates the neural correlates of cognition and their relationship with treatment outcomes in a large sample of older adults with treatment-resistant depression. The authors found that better cognitive function was associated with de-segregation of the DMN and FPN (lower functional connectivity between and within these networks), greater white matter integrity, and greater gray matter volume. Education and structural brain integrity were associated to better cognitive performance. Poorer cognitive function and reduced gray matter volume predicted continued treatment resistance.

Major Comments:

- Context and novelty: While it is valuable to characterize a sample of this size, many of the reported brain–cognition relationships (e.g., between structure and cognitive performance) seem consistent with previous findings in healthy aging and mild cognitive impairment (MCI) populations. As such, the novelty of the manuscript lies primarily in the predictive modeling of treatment outcome (MADRS remission). The finding that only structural MRI, but not DTI or rs-fMRI, predicts remission is noteworthy and merits deeper interpretation. Specifically, it would be helpful to discuss:
 - * Why structure, but not connectivity, predicts remission in late-life depression.
 - * Whether this reflects age-related processes (e.g., gray matter decline as a general vulnerability marker).
 - * How this aligns with or contrasts from previous findings, given that structural predictors have often shown limited predictive utility.
- Timing of imaging: The finding that gray matter structure predicts treatment outcome regardless of when imaging was performed (i.e., before or after treatment) deserves more discussion. This implies that the biomarker is predictive but not modifiable by treatment, which raises important questions about its clinical utility. The manuscript would benefit from explicitly addressing whether the authors view this as a prognostic rather than a treatment-sensitive marker, and what implications this might have for clinical stratification or intervention.
- Motion in rs-fMRI Analyses: The motion threshold set for the FC analyses is not extremely strict and it appears that the authors do not exclude high motion data points (scrubbing) ? Given the known sensitivity of functional connectivity measures to motion artifacts, could the authors:
 - * Report the distribution of motion across participants?
 - * Include motion as a covariate or test its impact on prediction? Given that other studies report FC to predict treatment outcome, this would be valuable to assess.
- Preregistration: Were the hypotheses for this specific paper of the OPTIMUM trial preregistered, and if so, where could one find these hypotheses? Why hypotheses were formulated for most outcomes, this was lacking for the DTI - it is only mentioned that lower WM volume is associated with poorer cognitive performance, but that's based on the volumetric data I think?
- White Matter Hyperintensities: The manuscript references white matter “intensities,” but it is unclear whether FLAIR imaging was included and analyzed. Given the known relevance of white matter hyperintensities (WMHs) in late-life depression, especially in relation to executive dysfunction, it would be good to describe FLAIR acquisition and analysis.

Minor/Specific Comments:

Introduction:

- * The initial abbreviation is slightly awkward, consider using “late-life treatment-resistant depression (LLD)”.
- * The statement that “remission may be protective from cognitive impairment” (line 77) is too strong; as shown in this study, the directionality could very well be reversed.

Results:

- * Table 1: Consider reporting non-parametric statistics for skewed variables (e.g., ATHF scores), especially if inclusion criteria imply non-normal distributions.
- * Please provide exact p-values for permutation tests where possible.
- * Lines 175–178 are long and somewhat repetitive; consider splitting and rephrasing for clarity.

Figures:

- * Figure 1D: Text suggests this is PLS1, but the figure refers to PLS2—please clarify.
- * Figure 1G: Not described in the legend—please update.
- * Figure 2G: Description in the text does not match the content—this needs to be checked.
- * Figure 3D: Only one PLS component's generalisability is shown. Could additional results be added to the supplement?

Supplementary Materials:

- * Suppl. Fig. 5: Use vendor names (Siemens, GE) rather than model names (Prisma, etc.), to avoid confusion.

(Remarks on code availability)

I have briefly scrolled through the code, and the scripts appear to be quite clear, even though they are not heavily commented. A read me is available. The code only contains the analysis following the processing of the data (freesurfer stats, FA values, FC estimates), so I think I should be able to run the code if I were to have data that were processed in the same way.

Reviewer #4

(Remarks to the Author)

In the manuscript “Brain-cognition relationships and treatment outcome in treatment-resistant late-life depression”, the authors investigate the relationship between cognitive functioning and neurobiological markers in older adults with treatment-resistant depression. Drawing on data from a multi-center randomized controlled trial, they examine baseline cognitive performance and neuroimaging measures including functional connectivity, white matter integrity, and gray matter structure in a sample of 397 participants, of whom 234 underwent MRI. Using multivariate statistical approaches, including partial least squares (PLS) and elastic net regression, the authors identify brain-cognition associations and attempt to predict treatment remission outcomes. They report that better cognitive function is associated with distinct patterns of brain connectivity and structure, and that poorer cognition and reduced gray matter volume are linked to continued treatment resistance. The study aims to elucidate the neural mechanisms underlying cognitive dysfunction in late-life depression and explore their prognostic value for treatment response.

While the study addresses a clinically meaningful and timely question using a rich dataset and advanced statistical techniques, I have serious doubts about the robustness and interpretability of the reported findings. In particular, the analytical approach appears to rely heavily on in-sample associations without sufficient attention to model validation and

generalizability. The use of PLS regression, a technique known to be prone to overfitting when not evaluated in a proper cross-validation framework, is central to many of the conclusions, yet there is no evidence that such validation was appropriately implemented. Moreover, the predictive modeling of treatment outcomes raises concerns about circularity, as it appears to build on features derived from analyses conducted on the full dataset. These issues call into question the stability and reproducibility of the reported associations and predictive effects. As such, I believe the manuscript requires substantial revision before it can be considered for publication.

PLS Analysis of Functional Connectivity and Cognition

Latent Variables and Model Specification:

The authors report two significant latent variables (LVs) from the PLS analysis linking functional connectivity to cognition, but it is unclear how the number of latent variables was decided upon. Was this choice guided by prior hypotheses or determined in a data-driven manner? If the latter, what criterion was used? Furthermore, if a second latent variable was extracted, was a deflation strategy applied to remove the variance explained by the first LV? These details are critical for interpreting the structure and independence of the reported LVs.

Generalizability and Overfitting:

The generalizability of the functional connectivity PLS model appears low and is not sufficiently acknowledged in the manuscript. Given that the reported results are likely to be overfitted, particularly in the absence of cross-validation, this limitation should be emphasized more explicitly. In general, PLS results derived without out-of-sample validation tend to be unstable and overly optimistic, and this concern applies equally to the analyses involving white matter and gray matter.

Reporting and Statistical Transparency:

There are inconsistencies in how p-values are reported throughout the manuscript. Some are given exactly, while others are expressed as thresholds (e.g., <0.05). I recommend reporting exact p-values in all cases, unless they are below 0.001. Additionally, the manuscript does not sufficiently describe the permutation test procedure used to assess the significance of the latent variables. What exactly was permuted in these tests? What were the p-values for LV1 and LV2? The method and rationale for associating individual functional connectivities or cognitive tests with the latent variables are also not clearly described. Was a statistical significance test applied to these associations, and if so, were multiple comparisons corrected for?

Discrepancies and Missing Correlations:

There is a notable inconsistency between the reported explained variance of 8.4 percent and the explained variance shown in Figure 2A. Moreover, the authors do not report the correlations between PLS1 Brain and PLS1 Cognition, or between PLS2 Brain and PLS2 Cognition. These values are fundamental to interpreting the strength and relevance of the latent variables. It is also unclear why the manuscript focuses so heavily on the first PLS mode while omitting a parallel discussion of the second.

Figure Presentation and Scatter Plots:

Figures 2D, 3D, and 4D contain scatter plots that are difficult to interpret and appear inconsistent across modalities. It is unclear what is being shown in each case. For example, the rationale for plotting the correlation between observed and predicted cognition is not explained, and it is not clear how the "observed cognition" scores were calculated. Typically, PLS results are illustrated by plotting the brain and cognition latent scores against one another to demonstrate the strength of the mode. The authors do not show these brain-cognition score correlations, which would be the appropriate way to assess model quality and generalizability. Additionally, all scatter plots shown appear to reflect in-sample fits. There is no cross-validated version of these plots or associated metrics, which further limits confidence in the robustness of the findings.

Predicting Treatment Outcomes

The section on predicting treatment remission raises several important concerns regarding the validity and transparency of the modeling approach. While the authors state that they tested their best-performing model in an out-of-sample fashion, only a single AUC value is reported, and it is unclear how this was derived. Although the methods section describes the use of cross-validation, the results do not reflect a proper cross-validation framework. There is no indication that performance was averaged across folds, nor are confidence intervals or variability measures reported. This lack of transparency makes it difficult to assess whether the results are truly generalizable.

Moreover, it is not clear whether features derived from the PLS analyses were included as predictors in the outcome model. If so, this would raise serious concerns about circularity or "double-dipping," unless the entire PLS procedure, including latent variable estimation and selection, was embedded within the cross-validation loop. As currently reported, there is no evidence that this was the case. Without such safeguards, the reported predictive performance is likely inflated and may not hold in independent samples. The authors need to clarify this modeling pipeline in full detail and, if necessary, re-run the analysis using a fully nested cross-validation framework that includes all preprocessing and dimensionality reduction steps.

General Conclusion

In light of the concerns outlined above, particularly regarding the lack of appropriate cross-validation, potential circularity in predictive modeling, and limited transparency in statistical procedures, I recommend that the manuscript undergo a major revision. The authors should re-implement their multivariate analyses and predictive models within a rigorous cross-validation framework that includes all preprocessing and dimensionality reduction steps. Permutation testing should be applied not to in-sample associations, but to the cross-validated model performance, in order to assess the statistical significance of generalizable patterns (<https://doi.org/10.1016/j.bpsc.2022.07.012>, <https://doi.org/10.1016/j.neubiorev.2020.10.030>). Addressing these issues would significantly strengthen the validity and

interpretability of the findings and improve the manuscript's contribution to the field.

(Remarks on code availability)

The code repository associated with this manuscript is unfortunately quite difficult to navigate and currently lacks sufficient documentation. Given the complexity of the analyses and the use of multiple neuroimaging modalities and statistical frameworks, a clearly structured and well-documented codebase is essential to ensure transparency and reproducibility. At present, it is challenging to trace how key results were generated or to verify the analytical pipeline as described in the manuscript. I strongly encourage the authors to substantially reorganize the repository, adopting a more intuitive folder structure and providing a comprehensive README file. This should include clear, step-by-step instructions on how to reproduce all major analyses and figures from the raw or preprocessed data. Improving the clarity and usability of the codebase will significantly enhance the value and credibility of the study.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

All my concerns have been addressed.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)

I only have one remaining recommendation, pertaining to the Results (comment #4). As requested, the authors reanalyzed the brain-cognition associations within each subgroup (cognitively unimpaired, MCI + dementia). For all models (rsFC, diffusion-weighted, and T1) the overall PLSR was non-significant in the subgroups (in contrast to the findings when the subgroups are combined). As the authors note, this suggests that the brain-cognition findings are driven by variability across the cognitive subgroups, and thus the models require the variability provided by including patients with various levels of impairment, and may not perform similarly if restricted to more homogenous groups alone (e.g., cognitively unimpaired alone). This is a critically important finding when considering generalizability and for implementation in future work. While these additional findings (and interpretation) were added to the Supplement, they should also be integrated into the primary manuscript to ensure these important results are emphasized to readers

(Remarks on code availability)

N/A

Reviewer #3

(Remarks to the Author)

I thank the authors for their revised work and answers to my questions. However, I still have some concerns regarding the timing of data acquisition (in line with the questions of reviewer 2 as well).

The argument that post-treatment MRI and cognitive measures can be treated as baseline predictors because baseline and 6-month values correlate strongly is not entirely convincing. High correlations primarily indicate rank-order stability across participants and do not rule out systematic treatment-related changes. Moreover, the statement that treatment would affect MRI and cognition in the same way is not well supported by the literature. Structural brain measures and cognitive outcomes may change on different timescales or show different sensitivities to treatment. Therefore, the justification for including measurements obtained after treatment onset as predictors of treatment response remains insufficient.

The additional subgroup analyses separating participants with and without MCI are appreciated. However, rerunning the PLS analyses in substantially smaller samples mainly results in a loss of statistical power, making it difficult to draw conclusions from the non-significant findings. The concern is more related to whether the brain-cognition relationships differ as a function of cognitive status. This question might be addressed more directly by incorporating MCI status as a grouping variable or covariate within the PLS framework, rather than splitting the sample. Such an approach would allow the authors to test whether the patterns differ between groups while retaining the full sample size.

(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/>

We thank the Reviewers for taking the time to provide helpful and insightful feedback on our submitted manuscript. Below we respond to each of the Reviewers' comments. Reviewer feedback is presented in blue, while our direct responses to the critiques are shown in *black italics*, while quoted text from the manuscript and the supplementary is shown in black, not italicized. Quoted text that we have now added or modified thanks to those critiques is also underlined. As a result of the Reviewers' comments, we believe that the manuscript has been substantially improved in terms of the quality of the analyses, results, and discussion.

Sincerely,

Aristotle Voineskos MD, PhD (on behalf of all authors)

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

The present manuscript addresses an important research question, examining how neural correlates of executive and memory functions relate to treatment outcomes in older adults with treatment-resistant depression. In a longitudinal study embedded in a multi-center randomized controlled trial, the authors assessed 397 participants at baseline with neuropsychological testing and obtained neuroimaging data (functional connectivity, diffusion metrics, and gray matter structure) in 234 of these. They applied multivariate partial least squares regressions to test associations between cognitive performance and brain measures, finding that 1) better cognition was linked to lower connectivity between default mode and frontoparietal networks as well as within the frontoparietal network; 2) reduced white matter tract integrity across several white matter tracts corresponded to poorer executive performance, and 3) gray matter structure was positively related to cognition. Finally, the authors used regularized elastic net logistic regression to identify baseline biomarkers predictive of remission and observed that poorer cognition and gray matter structure predicted treatment resistance.

I am impressed by the multimodal dataset and the rigor of the analyses: the validation steps (e.g., testing predictive performance of partial least squares or other models in held-out samples), integration of clinically relevant biomarker investigations, and comparison to null models to assess the added predictive value of neuroimaging variables all strengthen the study. Nonetheless, I have a few points for the authors to address.

We thank the Reviewer for their positive comments and for their constructive and helpful feedback. We aim to address the points raised in more detail below:

My major concern pertains to the sample size in predictive analyses: although recruiting older adults with treatment-resistant depression is inherently difficult and the cohort is impressive, the subset used to predict treatment outcomes in step 2 of the treatment protocol is barely 100 participants, which may inflate estimates of model performance (see, for example, <https://doi.org/10.1038/s41386-021-01020-7>). I recommend that the authors report the rates of treatment resistance in Table 1 to provide context for predictive prevalence and explicitly discuss the limited sample size and potential overfitting in the Limitations section.

We thank the Reviewer for raising this issue and have updated the Discussion as follows to acknowledge the sample size limitations and potential overfitting; we have also added the information to Table 1 as requested:

“Previous brain-cognition association studies in remitted LLD and MCI have found similar association levels between brain structure and cognitive function³³. While the performance of our remission prediction models was substantial (AUC>0.75), it can be further improved and needs to be tested in cross-trial analyses to establish generalizability^{53,54}. The sample size available for our neuroimaging models is a limitation. Previous work has shown that even for medium sample sizes like those in our study, models may overestimate predictive performance⁶⁰. Following prior recommendations, we aimed to ensure sizeable test sample sets, varying between 4-fold and 8-fold cross-validation. We also found clinical models ran in the smaller subset of participants with MRI and neurocognitive data to be highly consistent with clinical models ran in the larger subset of any participants who had clinical data.”

60. Flint C, Cearns M, Opel N, et al. Systematic misestimation of machine learning performance in neuroimaging studies of depression. *Neuropsychopharmacology*. 2021;46(8):1510-1517. doi:10.1038/s41386-021-01020-7

Finally, we have also added the remission rates to Table 1 as below:

Table 1

Demographic & Clinical Variables of Interest	OPTIMUM-NEURO Sample Characterization	
	Characteristics of Sample with Neuropsychological Data	Characteristics of Sample with both Neuropsychological and MRI data
N	397	234
Age (Years)	68.2 (5.9)	67.6 (5.4)
Female Sex (N)	268 68%	168 72%
Race (W/AA/A/NA)	354/28/8/7	209/16/4/5
Ethnicity (H/Non-H)	375/22	228/6
Education (Years)	14.8 (2.5)	14.7 (2.7)
MADRS score	19.5 (9.0)	19.7 (9.1)
<u>Remission rate Step 1</u>	<u>30.99%</u>	<u>32.77%</u>
<u>Remission rate Step 2</u>	<u>22.96%</u>	<u>25.88%</u>
Diagnosis (NC/MCI/DEM/NA)	178/197/13/9	116/105/6/7
<u>CDR (0/0.5/1/2/NA)</u>	<u>168/182/8/1/36</u>	<u>111/96/5/0/22</u>
ATHF score	8.0 (2.8)	8.2 (2.8)
CIRS-G score	8.3 (4.5)	8.9 (4.2)
MoCA score	24.6 (3.0)	24.9 (4.7)

In what may be of additional interest to readers, we have now also added clinical only models in the largest possible subset of the data to the Supplementary Section:

Supplementary Information

8. Clinical models predicting treatment outcomes

We compared the predictive model results in the sample of participants with MRI and clinical data to findings in the larger set of participants with clinical data, regardless of whether they also had MRI and neurocognitive data. We found consistent levels of performance in the full clinical sample. For step 1 (n=313), across folds AUC mean=0.65, SD=0.11. For step 2 (n=135), across all folds AUC mean=0.49.

SD=0.15. These values are consistent with the model performance observed in the subsample with neuropsychological data and in the subsample with both neuropsychological and MRI data (Figure 5).

Additionally, it will help readers if the Results clearly explain the different trial steps (e.g., steps 1 and 2) as treatment phases rather than analysis steps; currently this distinction only becomes apparent when inspecting the figure, which may confuse readers initially.

We thank the Reviewer for this helpful comment; we have clarified the step 1/ step 2 role as treatment arms as below:

Cross-validated elastic net regression models (Supplementary Information) showed varied prediction performance for remission ($\text{MADRS} \leq 10$) to bupropion or aripiprazole treatment in step 1 of the OPTIMUM clinical trial, and to lithium or nortriptyline in step 2 of the clinical trial, depending on the included predictors. The OPTIMUM trial included two steps, each 10 weeks' duration. In step 1, patients were randomly assigned 1:1:1 to a switch to bupropion or augmentation of their current antidepressant with bupropion or aripiprazole. In Step 2, patients who were ineligible for step 1, or who did not remit or otherwise benefit from their step 1 treatment, were randomly assigned 1:1 to a switch to nortriptyline or lithium augmentation. First, we found a cross-validation AUC of 0.64 when including cognitive data alongside baseline MADRS and demographics to predict remission to treatment offered in step 1 arm of the trial. In the sample with both cognitive and neuroimaging data the AUC for the same model was 0.66. This increased to an AUC of 0.74 when also including 74 brain structure variables (Figure 5A, 5B, 5C). When testing the best performing model including 16 cortical thickness variables, executive function and attention scores, and baseline MADRS as predictors, the out-of-sample AUC increased to 0.83 (specificity=0.70, sensitivity=0.78). Neither resting-state fMRI nor diffusion derivatives improved classification performance of remission in step 1 or step 2. When modeling treatment outcomes in step 2 arm of the trial, cortical thickness but not cognitive data was predictive of remission (Figure 5D, 5E, 5F). When testing the best performing model including three cortical thickness variables and baseline MADRS as predictors, the out-of-sample AUC increased to 0.78. Finally, we operationalized treatment response as change in MADRS scores, and found that a PLS regression with demographic, clinical, cognitive and brain structure data achieved considerable in-sample and hold-out accuracy (Figure 5G, 5H, 5I, 5J; PLS permutation $P < 0.001$) in predicting change in MADRS scores.

There are a few minor issues as well.

1. In the introduction, the definitions seem conflated: the first sentences refer to "Treatment-resistant depression (LLD) ..." but it is unclear whether "LLD" was intended to denote late-life depression or treatment-resistant late-life depression.

We agree with the Reviewer's point and have updated this throughout the manuscript. We refer to late-life depression as "LLD" and instead of abbreviating treatment-resistant LLD to "TRLLD" we use "treatment-resistant LLD" throughout the manuscript. We hope this way the manuscript is more accessible to a wide range of readers and doesn't introduce many new abbreviations.

"Treatment-resistant depression (LLD) is a chronic, debilitating disorder defined by the presence of persistent depressive symptoms despite two adequate trials of antidepressant medications. Treatment-resistance is particularly challenging in late-life depression (LLD), ..."

2. For the permutation tests of PLS, please provide exact p-values in the text, especially since many approach the significance threshold (e.g., around 0.03), so readers can judge robustness.

We thank the Reviewer for this suggestion and have updated the permutation p-values to exact values rather than a $p < 0.05$; we agree that this conveys more detailed information and is a more accurate report of our findings. We also updated the abstract to include specific values instead of a “ <0.05 ” threshold.

3. In Figure 2, the caption describes generalizability as “good,” but the main text characterizes it as “modest,” which seems more accurate given correlations between true and predicted scores are below $r=0.1$. Also, panels 2F and 2H appear to lack regression lines.

We agree with the Reviewer and have updated the figure legend as suggested:

“Further, we observed ~~good~~ modest generalizability of the results in held-out samples (D).”

4. Finally, specify the parcellation atlas used for gray matter analyses (e.g., if FreeSurfer’s Desikan–Killiany atlas plus subcortex), so readers can fully understand the anatomical definitions.

Following the Reviewer’s feedback, we have added the parcellation references in the methods:

“In region-wise analysis of brain structure, we included FreeSurfer-derived cortical thickness in the Desikan-Kiliany parcellation⁶⁸ and lateralized volumes of the hippocampus, amygdala and striatum corrected for total intracranial volume (aseg parcellation).”

Reviewer #2 (Remarks to the Author):

In a large cohort of older adults with treatment-resistant late life depression (LLD), the authors examined the neural correlates of cognitive function, and how these associate with prospective treatment outcomes following pharmacotherapy. The study has several notable strengths. This includes a large and robustly-phenotyped sample of older adults with treatment-resistant LLD, along with multi-modal neuroimaging data spanning structural and functional indices, and evaluation of both baseline and clinical trial data. Treatment-resistant late-life depression reflects a unique subset of older adults predisposed to poor medical outcomes and underrepresented in clinical research – which hinders treatment advancement. Studies in this population, asking questions like those posed here, are timely and necessary with potential for broad clinical impacts. However, there are several areas that require attention and further clarification to improve interpretation and impact. Please consider the following comments and suggestions which may help to strengthen the manuscript.

We thank the Reviewer for their helpful and insightful feedback; we have added several analyses and amended the manuscript to address these points below.

Key Points:

1. Specify which findings were evaluated cross-sectionally so readers can distinguish results derived from baseline vs interventional data. 2. Provide further detail regarding the ‘protocolized antidepressant treatment.’ If space does not permit a full description of the intervention, stating pharmacotherapy more broadly would aid interpretation so readers can distinguish between other antidepressant treatment modalities (e.g., behavioral interventions).

We thank the Reviewer for these suggestions which improve the informativeness of the key points section; we have added a section clarifying cross-sectional brain-cognition analyses, and have replaced the antidepressant treatment reference with pharmacotherapy in the key points (changes underlined):

Key points

“Findings In older adults with treatment-resistant LLD, higher connectivity between components of the default mode network and the frontoparietal network and within the frontoparietal network was associated with worse cognitive function across memory, executive function and language, while white matter circuit impairment was associated with worse executive function processing speed in cross-sectional analyses of baseline data. Lower cortical thickness and hippocampal volumes were also associated with worse memory and verbal fluency. Higher education duration was associated with better cognition. Both poorer cognitive performance and loss of gray matter predicted lower likelihood of remission to 10 weeks of pharmacotherapy protocolized antidepressant treatment administered as part of the clinical trial.”

Abstract:

1. Please report the cognitive status of participants (e.g., unimpaired, MCI).

We have added the following based on the Reviewer feedback:

Abstract:

In the context of a longitudinal biomarker study embedded within a multi-center randomized controlled trial for late-life treatment-resistant depression, 397 participants completed baseline neuropsychological testing, and of these 234 adults (117 no cognitive disorder, 102 MCI, 6 with mild dementia) successfully completed a baseline MRI scan.

Methods:

“Cognitive status (no cognitive disorder, MCI, and dementia diagnoses) was determined via adjudication following DSM-5 and NIA-AA 2011 criteria with a multi-disciplinary team of clinical research staff, neuropsychologists, and psychiatrists based on the cognitive assessment scores and a wide range of other relevant clinical factors, including, medical history, occupational and educational attainment, etc.”

2. Report the type of intervention participants received.

Similar to the language we adopted in the key points section following Reviewer 2’s comment, we refer to pharmacotherapy in the abstract and state in the methods each of the drugs administered:

Abstract

“Further, we employed regularized elastic net regressions to identify biomarkers predicting depression remission ($MADRS \leq 10$) to pharmacotherapy in the clinical trial.”

Methods section included in the original submission:

“The parent trial included two steps, each 10 weeks’ duration. In step 1, patients were randomly assigned 1:1:1 to a switch to bupropion or augmentation of their current antidepressant with bupropion or aripiprazole. In Step 2, patients who were ineligible for step 1, or who did not remit or otherwise benefit from their step 1 treatment, were randomly assigned 1:1 to a switch to nortriptyline or lithium augmentation.”

3. “Education and better structural brain maintenance, but not overall health were associated with better cognition”. It’s not clear what structural brain maintenance reflects or how this metric is defined. The index for overall health is also not reported. Please specify.

While the Methods and Results detail the structural brain maintenance measure (normative deviation in brain structure from a sample of ~100k participants in Bethlehem et al) and overall health (CIRS-G), these aren’t critical to our overall conclusions and we removed the reference to these two findings from the abstract to focus on the main findings:

Greater education ~~and better structural health but not overall health were also~~ associated with better cognition.

Introduction:

1. Consistency of population and brain network descriptors. Treatment-resistant LLD and LLD are used interchangeably throughout, though as the authors correctly note, treatment resistant LLD is associated with distinct clinical trajectories. They also switch between two descriptions for the same network (frontoparietal and executive control). Please select one for consistency and ease of understanding.

Relatedly, the authors use 'LLD' as the acronym for treatment resistant depression in sentence 1 of the introduction.

We thank the Reviewer for pointing out this issue, we agree the first instance of LLD abbreviation is incorrect and have amended this for LLD to refer to late life depression and we explicitly say 'treatment resistant LLD' in contexts discussing our sample or others investigating treatment resistant LLD.

Treatment-resistant depression (LLD) is a chronic, debilitating disorder defined by the presence of persistent depressive symptoms despite two adequate trials of antidepressant medications. Treatment-resistance is particularly challenging in late-life depression (LLD), as less than one-third of patients with treatment-resistant LLD remit, even in augmentation treatment trials¹⁻³.

We now also use frontoparietal networks throughout the manuscript following the Reviewer's suggestion, e.g.:

"We hypothesized that loss of gray and white matter alongside with de-segregation of the executive control, frontoparietal and cortico-limbic circuits will be associated with worse cognitive performance in cross-sectional analyses."

Methods:

1. Methods state "Exclusion criteria included severe neuropsychiatric conditions such as Parkinson's Disease or schizophrenia." Parkinson's is typically classified as neurodegenerative rather than neuropsychiatric condition. In addition, were those with neurodegenerative conditions also excluded at study start?

2. Relatedly, the cognitive status of participants is difficult to ascertain. The methods state that those with dementia (Short Blessed Test ≥ 10) were excluded. Yet, Figure 1 reports that the sample included subjects with various levels of cognitive function and Table 1 reports that there were 13 subjects with probable dementia included in the sample. Further clarification on the inclusion/exclusion and cognitive status of participants is needed.

We thank the Reviewer for bringing this to our attention and clarify in the methods that patients with neurodegenerative diagnoses were also excluded:

"Exclusion criteria included severe neuropsychiatric and neurodegenerative conditions such as..."

Patients with significant cognitive impairment at the start of the study were excluded via the $SBT \geq 10$ threshold as stated in the current version of the methods. As is the case with most if not all brief dementia screening tools, using a single Short Blessed Test (SBT) threshold or cutoff score will result in exclusion of some participants with lifelong relatively low cognitive function who do not have dementia and most relevant to this discussion, inclusion of some lifelong high functioning individuals in the early stages of a neurodegenerative disorder. The in-depth ADRC-style cognitive diagnostic adjudication process which was conducted once participants were active in the study entailed careful review of a substantial amount of clinically relevant data, accounting for a wide range of factors that influence cognition, resulting in more refined and valid diagnoses. Thus, it is not surprising that a small proportion of our participants, who entered the study with $SBT > 10$, were ultimately diagnosed with (mild) dementia.

We report adjudication diagnoses in this manuscript given that these are based on ADRC-style multi-disciplinary adjudications that consider performance on a range of cognitive tests; and a wide range of other relevant clinical factors, including, medical history, occupational and educational attainment, etc. These diagnoses generally map well onto CDR scores, where CDR-0 = cognitively normal, CDR-0.5 is

MCI, and CDR-1=dementia. We now also include CDR scores in the Table 1 next to adjudication diagnoses given these are commonly used in neurology studies and may add additional context to the adjudication diagnoses.

Table 1:

Demographic & Clinical Variables of Interest	OPTIMUM-NEURO Sample Characterization	
	Characteristics of Sample with Neuropsychological Data	Characteristics of Sample with both Neuropsychological and MRI data
N	397	234
Age (Years)	68.2 (5.9)	67.6 (5.4)
Female Sex (N)	268 68%	168 72%
Race (W/AA/A/NA)	354/28/8/7	209/16/4/5
Ethnicity (H/Non-H)	375/22	228/6
Education (Years)	14.8 (2.5)	14.7 (2.7)
MADRS score	19.5 (9.0)	19.7 (9.1)
<u>Remission rate Step 1</u>	<u>31.0%</u>	<u>32.8%</u>
<u>Remission rate Step 2</u>	<u>23.0%</u>	<u>25.9%</u>
Diagnosis (NC/MCI/DEM/NA)	178/197/13/9	116/105/6/7
<u>CDR (0/0.5/1/2/NA)</u>	<u>168/182/8/1/36</u>	<u>111/96/5/0/22</u>
ATHF score	8.0 (2.8)	8.2 (2.8)
CIRS-G score	8.3 (4.5)	8.9 (4.2)
MoCA score	24.6 (3.0)	24.9 (4.7)

We also note that many patients with late life depression exhibit cognitive impairment, citing the below in the introduction:

“In addition, high rates of cognitive impairment in this population^{4,5} significantly increase the risk of dementia⁶,”

4. Bhalla RK, Butters MA, Becker JT, et al. Patterns of Mild Cognitive Impairment After Treatment of Depression in the Elderly. *Am J Geriatr Psychiatry*. 2009;17(4):308. doi:10.1097/JGP.0B013E318190B8D8
5. Butters MA, Whyte EM, Nebes RD, et al. The nature and determinants of neuropsychological functioning in late-life depression. *Arch Gen Psychiatry*. 2004;61(6):587-595. doi:10.1001/archpsyc.61.6.587

3. According to the supplement, there were 212 subs with useable T1 data, but 219 subs with useable DTI data included in the analyses. Often, T1 data is used in the preprocessing of DTI data. It's unclear how this would result in more participants with DTI data included in the analyses than T1, unless T1 wasn't used in the preprocessing. Please clarify.

We used more stringent QC thresholds for cortical thickness quantification than for using T1s in registration of the DTI data. Specifically, the DTI processing stream is using streamline registration from Slicer

https://github.com/SlicerDMRI/whitematteranalysis/blob/master/bin/wm_register_to_atlas_new.py) which is independent from Freesurfer. We have visually QC'd the tract bundles for each person independently from Freesurfer, and have applied more stringent criteria for FreeSurfer cortical mantle reconstruction resulting in more DTI datapoints than FreeSurfer datapoints being included.

4. "Among 73 reconstructed tracts, we excluded tracts with more than 3% unusable data and imputed missing data in the remaining 62 tracts." How was imputation performed?

Following Reviewer feedback we now clarify that we used mean imputation:

"Among 73 reconstructed tracts, we excluded tracts with more than 3% unusable data and imputed missing data in the remaining 62 tracts using mean imputation."

Results:

1. It appears that about 1/3 of the sample underwent cognitive testing and neuroimaging after treatment started (per Supplemental Table 3). To address this, the authors conducted sensitivity analyses excluding participants with neuropsychological testing completed over six weeks after the start of treatment. Can the authors provide justification for selecting 6 weeks after treatment initiation?

We thank the Reviewer for highlighting this – we have added the following justification to supplementary:

"Previous data show that 4 weeks may not be sufficient to assess efficacy in non-treatment resistant depression¹. In addition the Antidepressant Treatment Response Questionnaire requires an adequate monotherapy trial to be at least 6 weeks long²."

1. Quitkin FM, Rabkin JG, Ross D, Mcgrath PJ. Duration of Antidepressant Drug Treatment: What Is an Adequate Trial? *Arch Gen Psychiatry*. 1984;41(3):238-245. doi:10.1001/archpsyc.1984.01790140028003

2. Chandler GM, Iosifescu D V., Pollack MH, Targum SD, Fava M. RESEARCH: Validation of the Massachusetts General Hospital Antidepressant Treatment History Questionnaire (ATRO). *CNS Neurosci Ther*. 2010;16(5):322. doi:10.1111/J.1755-5949.2009.00102.X

2. Related to the above, did the authors run sensitivity analyses focusing only on subjects whose cognitive/neuroimaging data was collected before treatment began? Treatment duration was 10 weeks. If the aim is to examine cognitive/MRI predictors of treatment response, and the cognitive/MRI predictors are collected after treatment begins, this could significantly impact the findings and what those findings mean. Prior studies have shown changes in resting state connectivity and even modest changes in cognition following pharmacotherapy of brief duration (1-4 weeks). As the authors acknowledge in the discussion, it's possible treatment impacted brain and cognitive markers. Demonstrating that the results persist in the subset of those whose predictors (cognitive performance and brain MRI) were collected before treatment commenced would significantly strengthen the results and interpretation.

We thank the Reviewer for raising this important issue. There are two sets of analyses in this paper that may be affected by the fact that >1/3 of the sample had their MRI and cognitive data collected after the start of treatment.

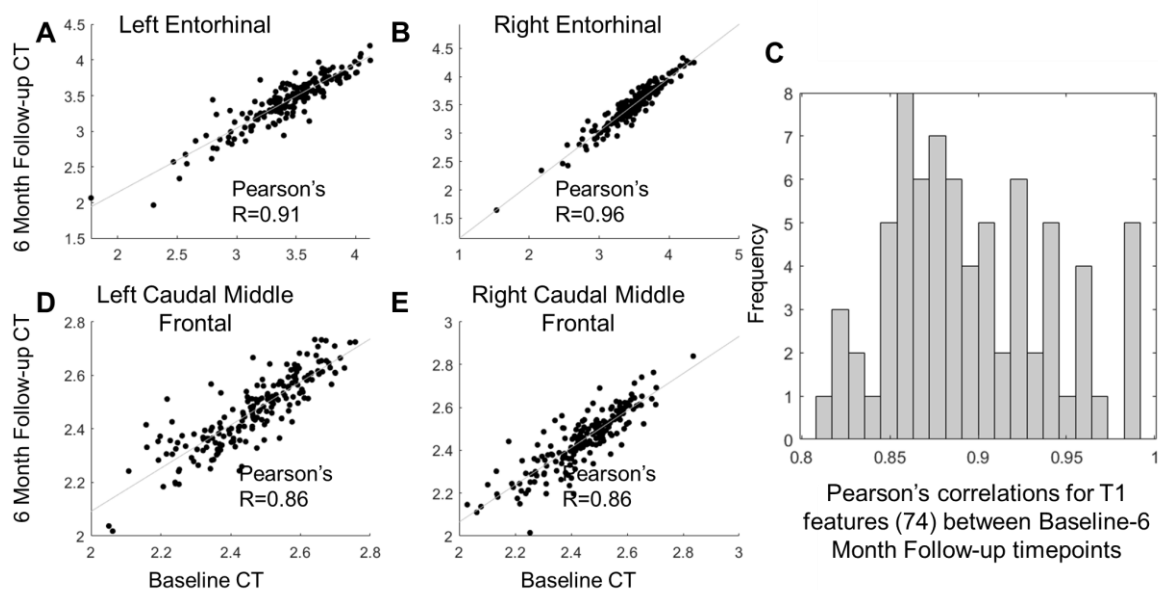
First, the predictive model utilizes 'baseline' MRI and cognition to predict remission to treatment and is more likely to be affected. Our best predictive models include structural MRI and some cognitive variables. From prior work we know that structural changes usually occur over years rather than over months; data from the OPTIMUM-NEURO participants analyzed here supports this; specifically, we add a supplementary section (quoted below) showing high consistency between baseline and 6-month follow-up structural MRI metrics and relatively high consistency for cognitive tests. Given that our predictive models found structural MRI markers such as those shown below and cognitive variables to predict remission, we argue that this consistency justifies the use of all the data collected in predictive modeling. We are hesitant to exclude participants from these modeling analyses given that it would decrease the sample sizes from this clinically ill and rarely studied population.

Second, we conducted brain-cognition association analyses in this manuscript. However, MRI and cognition was collected in close succession (either on the same day or within a few days from each other). Therefore, treatment would affect both MRI and cognition in the same way. We have amended the Discussion to reflect this argument as cited below.

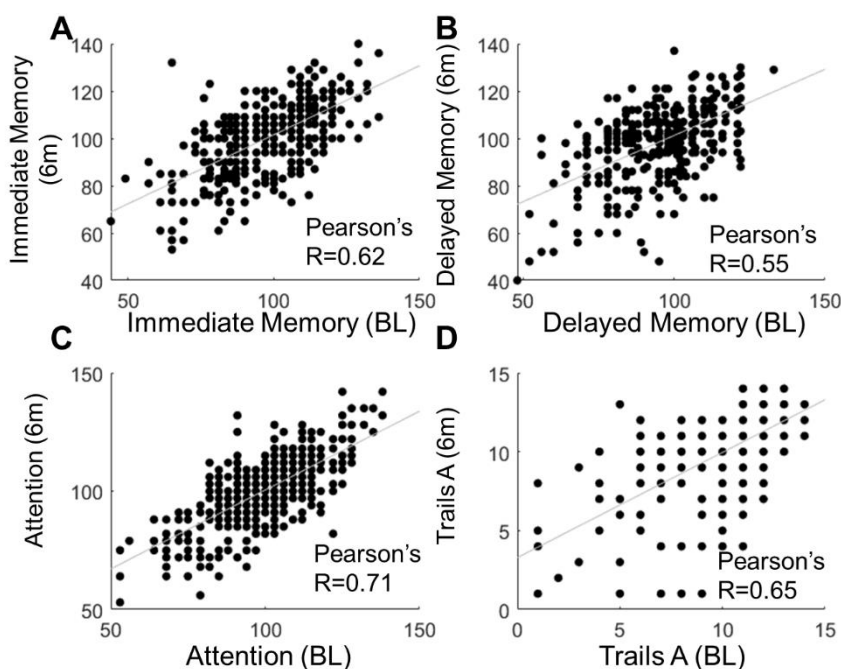
Supplementary Information

9. Stability of structural MRI and cognitive function between baseline and 6 month follow up

To provide a justification for using structural MRI features that were potentially collected several weeks after the start of the clinical trial (Supplementary Figure 2) in predicting prediction clinical trial outcomes, we leverage the 6-month follow-up data from OPTIMUM-NEURO. Specifically, we correlate baseline cortical thickness and subcortical volume values with those collected at 6 months, after longitudinal FreeSurfer processing on top of fMRIPrep. We found remarkable consistency of baseline-6 month correlations across all structural features investigated (Supplementary Figure 7), with low within-subject variability across the 6-month follow-up period. This data supports the use of structural features ascertained even several months after the start of the clinical trial. Further, we found good consistency in cognitive performance between baseline and 6-month timepoints (Supplementary Figure 8), supporting the use of cognitive tests in predicting remission.



Supplementary Figure 7. Baseline-to-6 month follow-up correlations between structural features. Left and right entorhinal (A, B), and left and right caudal middle frontal gyri (D,E) were chosen as representative examples from medial temporal lobe and the prefrontal cortex. Distribution of all baseline-6 month correlations is shown in (C), with values ranging between 0.8 and 0.99. CT: cortical thickness.



Supplementary Figure 8. Baseline-to-6 month follow-up correlations between key cognitive tests used in the analyses. Immediate and delayed memory (A, B), attention (C) and Trailmaking A (panel D) showed a high degree of consistency between baseline and 6 month. All $p<0.0001$.

It is possible that fMRI and DTI data may not be this stable over the period of several months, which we acknowledge in the limitations:

Discussion:

"Further, although cognitive function and brain MRI were assessed at varying time intervals relative to the OPTIMUM treatment, our supplementary data show that brain structure and cognition were remarkably stable over a 6-month time period. It is possible that OPTIMUM treatment has impacted functional and diffusion MRI, and we found these markers were not predictive of remission. Future studies using a more controlled design may therefore identify connectivity markers of remission in LLD. Both brain and cognitive data were collected at the same interval relative to the OPTIMUM treatment, supporting out cross-sectional brain cognition findings. Our supplementary results support the former hypothesis, given that excluding participants with post-treatment scans and cognitive testing did not substantially change the results."

3. Given that the study is focusing, in part, on predictors of treatment response, can the authors report in the results the proportion of responders in Step 1 and Step 2?

Following Reviewer feedback we have added remission rates to Table 1, citing Table 1 in response to Reviewer's point above.

4. More than 40% of the sample was classified as having MCI. A core aim of the manuscript is to evaluate the neural correlates of cognitive function, and it is very likely that these relationships would differ based on cognitive status, given structural and functional brain changes linked to MCI. It is recommended that the authors examine whether the pattern and directionality of the findings differ based on cognitive status (unimpaired vs. MCI).

Following Reviewer's suggestion, we have reanalyzed the brain-cognition associations in the no cognitive disorder/unimpaired and cognitive impairment subgroups and added discussion of the findings to the supplementary:

Supplementary information:

10. Brain-cognition associations in subsets of participants

We next tested whether our findings in the overall OPTIMUM sample hold up in subsets of participants with MCI or dementia, and in those with no cognitive disorder. For the resting-state FC models, we found that in each subgroup individually, the overall PLSR was not significant, i.e. did not explain more variance in cognitive function compared to chance permutation models (cognitively normal subjects: n=112, $p_{\text{PERMUTATION}}=0.74$ for a 3-latent variable model; MCI and dementia subgroup: n=97, $p_{\text{PERMUTATION}}=0.36$ for a 3-latent variable model). For the diffusion-weighted fractional anisotropy models, we found that in each subgroup individually, the overall PLSR was not significant (cognitively normal subjects: n=109, $p_{\text{PERMUTATION}}=0.30$ for a 1-latent variable model; MCI and dementia subgroup: n=103, $p_{\text{PERMUTATION}}=0.12$ for a 1-latent variable model). For the T1 cortical thickness and subcortical volume models, we found that in each subgroup individually, the overall PLSR was not significant (cognitively normal subjects: n=113, $p_{\text{PERMUTATION}}=0.28$ for a 3-latent variable model; MCI and dementia subgroup: n=103, $p_{\text{PERMUTATION}}=0.54$ for a 3-latent variable model). This suggests that our brain-cognition findings are driven by the variability in cognitive function across the two patient subgroups. Greater variability in brain data and cognitive/behavioral performance likely drives larger effect sizes⁴ and leads to significant findings in the whole sample. Reducing variability by restricting the sample to the more homogenous subgroups removes this effect.

We have also added CDR scores to add information that may be useful in addition to the adjudication diagnosis.

5. "Similarly, lower connectivity of different FPN components with each other was also robustly associated with cognitive performance." Specify the directionality (better or worse cognitive performance).

We thank the Reviewer for pointing out this section – we have updated it to be more specific:

~~"Similarly, Lower connectivity of left and right FPN components with each other~~ was also robustly associated with better cognitive performance."

6. Part A of the Trail Making Test is described throughout the manuscript as an executive function test. However, Trails A is widely considered to be a measure of visual scanning and psychomotor/processing speed. Trail Making Test Part B, in contrast, is classified as an executive function measure. The description of Trails A and interpretation of the various results pertaining to this measure should be updated throughout to reflect this.

We thank the Reviewer for pointing this out. We agree that Trails A is a mix of psychomotor and processing speed rather than executive function, and we have replaced all mention of executive

function in this context throughout the manuscript, starting with key points, abstract, results and methods.

7. The abstract mentions results pertaining to ‘overall health,’ but these findings do not seem to be included in the results.

We have removed this mention from the abstract; we had explored the CIRS-G as a measure of overall health but decided not to include it in the analyses to keep them focused and given the limitations of CIRS-G as a measure of overall health. Instead, we present overall CIRS-G scores for descriptive purpose in table 1 only.

8. “When operationalizing treatment response as change in MADRS scores, a PLS regression with demographic, clinical, cognitive and brain structure data achieved considerable in-sample and hold-out accuracy.” Please specify what clinical and demographic data were included. For instance, were sex and age incorporated?

Indeed, the baseline clinical model included age, sex, and baseline MADRS; we clarify this in addition in the methods:

“The clinical model included age, sex, and baseline MADRS scores.” Page 16, regarding the response/nonresponse models

“In addition, we have used a PLS model predicting MADRS change (absolute difference in MADRS score between baseline and the end of last step completed) from clinical (MADRS), demographic (age, sex), cognitive and brain structure data.” Page 16, regarding PLS

Discussion:

1. The AUC’s reported are generally below the thresholds considered clinically useful. This should be acknowledged in the discussion.

While we agree that remission prediction performance can be improved, an AUC>0.75 and balanced accuracy of >75% for predicting remission is quite substantial and indeed several clinical trials are currently underway to test predictors with similar levels of performance (NCT05017311, NCT05537584, NCT06408246, NCT04181736, and others) in a prospective RCT setting. We therefore qualify the discussion as below:

“While the performance of our remission prediction models was substantial (AUC>0.75), it can be further improved and needs to be tested in cross-trial analyses to establish generalizability^{53,54}.”

2. It appears that white matter (FA) explained much less variance relative to resting state functional connectivity (1.6% WM vs. ~8% with rsFC). Can the authors speculate about why this may be the case? In addition, the Trail Making Test, Part A, was the only cognitive test related to FA. Whereas resting state FC was related to 12 cognitive tests comprising multiple domains. The authors may consider discussing the differences between the white matter and connectivity findings and what this may suggest regarding the neural correlates of cognitive dysfunction in treat-resistant LLD.

We thank the Reviewer for this suggestion and have revised the Discussion in response. As the Reviewer pointed out, for the WM analysis, the only cognitive test associated with brain data was the Trail Making A, whereas FC was related to several tests across domains. Since the % variance

explained metric refers to the amount of variance in all Y scores that the PLS can explain, it is lower when the associations we are seeing are specific to just one of the 12 included cognitive variables. We have previously included language acknowledging this in the Discussion and have refined this section as below:

Discussion:

“We also found that lower white matter integrity of several tracts (e.g superior longitudinal fasciculus) was specifically associated with worse Trail Making A performance. This finding is consistent with previous studies showing that reduced integrity of tracts connecting prefrontal and parietal areas, including the superior longitudinal fasciculus, are associated with worse processing speed executive function in healthy individuals³², in those with varying levels of cognitive impairment³³, but also in children³⁴ and young adults³⁵. Similarly, faster processing speed is associated with higher FA in tracts including corpus callosum^{36,37} and average cerebral FA³⁸, and cortico-striatal tracts³⁹. Unlike gray matter and functional connectivity, FA was not significantly associated with episodic memory performance in our analysis, consistent with previous studies of white matter^{39,40}. We also show that white matter hyperintensities, an index of cerebrovascular disease burden, were predictive of both cognitive function and FA scores, consistent with the theory that cerebrovascular pathology affecting white matter may be driving cognitive impairment⁷. Our findings further support the hypothesis that initial impairment in white matter in treatment-resistant LLD may predominantly impact processing speed rather than memory and language domains, while functional brain organization is associated with cognitive function more broadly.”

3. In Step 1, both cortical thickness and cognition were predictive of remission. In step 2, cortical thickness but not cognitive data predicted remission. Do the authors have theories as to why this might be the case? Could it have something to do with the drug class or characteristics of responders to certain pharmacotherapies but not others?

We thank the Reviewer for this suggestion and have added the following to the discussion:

Discussion

“Cognitive performance predicted remission to bupropion or aripiprazole in step 1, but not to lithium or nortriptyline in step 2, suggesting that different markers may be predictive for remission to distinct drug classes. Patients in step 2 arguably were more treatment-resistant than those in step 1, with lower step 2 remission rates (~20% vs ~30%), suggesting somewhat distinct biological mechanisms and predictive biomarkers may be present in these LLD subgroups.”

4. “Cognitive performance and gray matter integrity predicted remission after 10 weeks of bupropion or aripiprazole treatment in the OPTIMUM trial.” The authors may consider rephrasing this for accuracy, as a large portion of participants completed cognitive/MRI assessments after drug treatment started (i.e., for many subjects, it was less than 10 weeks of treatment duration after cognitive/MRI data collection).

Following Reviewer’s suggestion, we have revised this sentence:

“Cognitive performance and gray matter integrity predicted remission ~~after 10 weeks of~~ to bupropion or aripiprazole treatment in the OPTIMUM trial.”

5. “Our supplementary results support the former hypothesis, given that excluding participants with

post-treatment scans and cognitive testing did not substantially change the results.” Based on the supplement, it sounds like sensitivity analyses excluded those whose cognitive/MRI data was obtained 6 weeks after treatment began, as opposed to excluding all subjects whose cognitive/MRI data was obtained after treatment started. Please clarify.

We thank the Reviewer for highlighting this issue – we have now included information on the stability of baseline-6 month data in the supplementary information as cited in response to the Reviewer’s point 2 raised in the Results section above. We have also removed the sentence regarding sensitivity analyses from the Discussion.

Reviewer #3 (Remarks to the Author):

This study investigates the neural correlates of cognition and their relationship with treatment outcomes in a large sample of older adults with treatment-resistant depression. The authors found that better cognitive function was associated with de-segregation of the DMN and FPN (lower functional connectivity between and within these networks), greater white matter integrity, and greater gray matter volume. Education and structural brain integrity were associated to better cognitive performance. Poorer cognitive function and reduced gray matter volume predicted continued treatment resistance.

We thank the Reviewer for the helpful and constructive feedback and have aimed to address each point raised below. We agree with the Reviewer's points and have added several Supplementary sections to address the issues raised. We hope these improve the manuscript and strengthen our results and conclusions.

Major Comments:

- Context and novelty: While it is valuable to characterize a sample of this size, many of the reported brain–cognition relationships (e.g., between structure and cognitive performance) seem consistent with previous findings in healthy aging and mild cognitive impairment (MCI) populations. As such, the novelty of the manuscript lies primarily in the predictive modeling of treatment outcome (MADRS remission). The finding that only structural MRI, but not DTI or rs-fMRI, predicts remission is noteworthy and merits deeper interpretation. Specifically, it would be helpful to discuss:

* Whether this reflects age-related processes (e.g., gray matter decline as a general vulnerability marker).

* Why structure, but not connectivity, predicts remission in late-life depression.

* How this aligns with or contrasts from previous findings, given that structural predictors have often shown limited predictive utility.

We agree that our brain-cognition findings are quite consistent with existing MCI data, and highlight this in the final conclusions section, highlighting the novelty of the LLD sample.

“In conclusion, varying levels of cognitive performance in older adults with treatment-resistant LLD have distinct neural correlates, with protective effects of education and structural brain maintenance. Our findings are consistent with existing literature in healthy aging and MCI, suggesting that many brain-cognition findings such as the role of DMN and FPN in memory and executive function, or protective effects of education are conserved in treatment-resistant LLD. Our novel predictive models are promising and, once tailored to individual drugs, should be tested prospectively (1UG3MH137353-01)....”

Age was included as a predictor in the elastic net models and did not strongly contribute to the model predictions so we believe the gray matter finding is more reflective of a brain health/MCI related predictive signal:

“In older adults with considerable variability in gray matter integrity and cognitive performance, we found gray matter structure but not brain function to predict remission, potentially suggesting distinct biomarkers in TRLDD and adult MDD. This contribution of gray matter integrity is independent from age, as age was included in the models but was not predictive.”

Regarding the Reviewer's point about structural vs functional predictors, there is a wide range of literature showing both functional and structural MRI predictors of prognosis and treatment outcome (e.g. Viviano JD, Buchanan RW, Calarco N, et al. Resting-State Connectivity Biomarkers of Cognitive Performance and Social Function in Individuals With Schizophrenia Spectrum Disorder and Healthy

Control Subjects. Biol Psychiatry. 2018;84(9):665-674. doi:10.1016/J.BIOPSYCH.2018.03.013; Koutsouleris N, Kambeitz-Ilankovic L, Ruhrmann S, et al. Prediction Models of Functional Outcomes for Individuals in the Clinical High-Risk State for Psychosis or With Recent-Onset Depression: A Multimodal, Multisite Machine Learning Analysis. JAMA Psychiatry. 2018;75(11):1156-1172. doi:10.1001/JAMAPSYCHIATRY.2018.2165 but also Zhukovsky et al JAMA Network Open). What is unique about our study is the rigorous clinical trial during which the neuroimaging was conducted, which allowed for detailed information of patient characteristics, and even more importantly medication type, dose, duration of administration, and response (or non-response) to treatment. Our participants are older adults who show a wide range of cognitive performance and loss of gray matter, contrasting with most adult studies of MDD in whom brain structure is preserved. This provides a wider range/more variance in the predictor, which statistical simulations have shown usually helps discover robust effects. Ultimately these findings will be tested again in an ongoing clinical trial that our collaborative team is leading, which will reveal in a prospective manner whether the use of neuroimaging to determine medication treatment assignment leads to better outcomes for patients with treatment resistant LLD (compared to usual care). This will be the true 'litmus test' of this neuroimaging marker.

- Timing of imaging: The finding that gray matter structure predicts treatment outcome regardless of when imaging was performed (i.e., before or after treatment) deserves more discussion. This implies that the biomarker is predictive but not modifiable by treatment, which raises important questions about its clinical utility. The manuscript would benefit from explicitly addressing whether the authors view this as a prognostic rather than a treatment-sensitive marker, and what implications this might have for clinical stratification or intervention.

We thank the Reviewer for raising this point. We agree that structural MRI markers are prognostic and not modifiable by treatment. We have added this info in the key points:

Key points:

"We also identify potential neural and cognitive prognostic markers of ongoing treatment resistance."

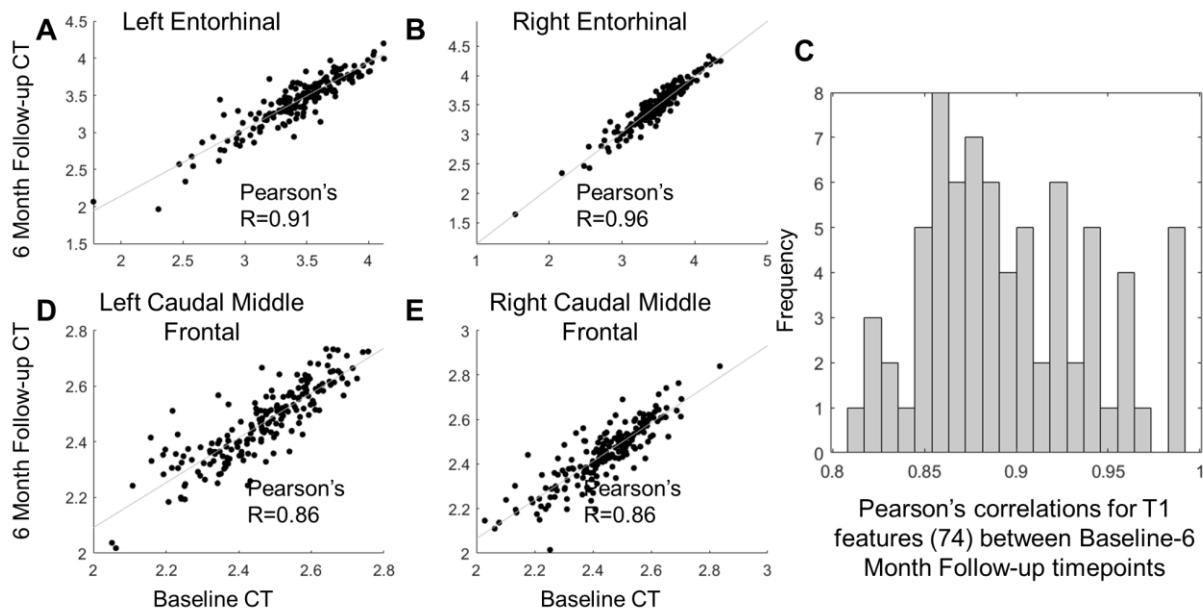
And also in the Discussion:

"The structural MRI markers predicting remission are remarkably stable over the span of a few months, suggesting this marker may have prognostic potential, which can be tested directly in future studies."

We have also added the below preliminary data to support the use of structural markers to the Supplementary:

9. Stability of structural MRI and cognitive function between baseline and 6 month follow up

To provide a justification for using structural MRI features that were potentially collected several weeks after the start of the clinical trial (Supplementary Figure 2) in predicting prediction clinical trial outcomes, we leverage the 6-month follow-up data from OPTIMUM-NEURO. Specifically, we correlate baseline cortical thickness and subcortical volume values with those collected at 6 months, after longitudinal FreeSurfer processing on top of fMRIPrep. We found remarkable consistency of baseline-6 month correlations across all structural features investigated (Supplementary Figure 7), with low within-subject variability across the 6-month follow-up period. This data supports the use of structural features ascertained even several months after the start of the clinical trial. Further, we found good consistency in cognitive performance between baseline and 6-month timepoints (Supplementary Figure 8), supporting the use of cognitive tests in predicting remission.



Supplementary Figure 7. Baseline-to-6 month follow-up correlations between structural features. Left and right entorhinal (A, B), and left and right caudal middle frontal gyri (D,E) were chosen as representative examples from medial temporal lobe and the prefrontal cortex. Distribution of all baseline-6 month correlations is shown in (C), with values ranging between 0.8 and 0.99. CT: cortical thickness.

We also include the baseline-6 month data in supplementary section 9, and cite it in response to Reviewer 2 above, alongside a more extended discussion of the impact of the timing of the biological assessments relative to the trial. We did not reproduce the entire Supplementary Section 9 here in order to keep the response letter a reasonable length.

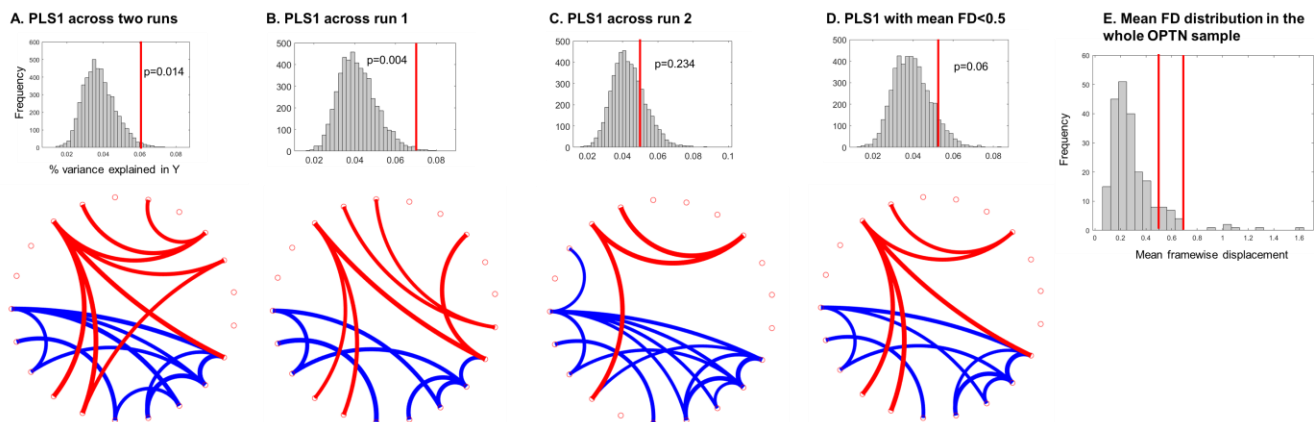
- Motion in rs-fMRI Analyses: The motion threshold set for the FC analyses is not extremely strict and it appears that the authors do not exclude high motion data points (scrubbing) ? Given the known sensitivity of functional connectivity measures to motion artifacts, could the authors:

* Report the distribution of motion across participants?

* Include motion as a covariate or test its impact on prediction? Given that other studies report FC to predict treatment outcome, this would be valuable to assess.

We agree that motion is an important confound and we have indeed included motion as one of the confound variables alongside age and sex in the original analyses. Rather than directly including motion in the PLSR model we have used ComBat, harmonizing for mean framewise displacement. We detail this in the methods section and regressing out mean FD from the fmri data using ComBat is a part of our processing pipeline. We did not scrub individual timeseries given the field does not agree on whether this is the best approach. While the XCP toolbox implements scrubbing, there are concerns that scrubbing potentially removes useful signal and by breaking up the BOLD timeseries introduced additional artifacts. UK Biobank and FSL pipelines do not include scrubbing or global signal regression by default, hence we have not included these steps. We did exclude high-motion subjects however. Specifically, we excluded participants with mean FD >0.7 in main analyses and applied a mean FD threshold of 0.5 in Supplementary:

“Further, we conducted another PLS analysis excluding participants with mean FD>0.5. This PLS model did not explain more variance than expected by chance ($p=0.06$); however, bootstrapping revealed that many of the connectivities captured by the main model were also captured in this PLS model ($|Z|>3$, Figure S4D).



Supplementary Figure 4. Sensitivity analyses showing FC-cognition associations for each of the two runs of rs-fMRI data (B,C) acquired and with a stricter FD threshold of 0.5 (D) alongside the main analysis (A). Motion distribution across all participants, including the 0.6 and 0.7 thresholds (E).

We have now also added a distribution plot of the mean FD variable to the supplementary Figure S4E.

- Preregistration: Were the hypotheses for this specific paper of the OPTIMUM trial preregistered, and if so, where could one find these hypotheses? Why hypotheses were formulated for most outcomes, this was lacking for the DTI - it is only mentioned that lower WM volume is associated with poorer cognitive performance, but that's based on the volumetric data I think?

We pre-registered the study and the main aims and analyses are described here: NCT05204940. The overall focus and main aims of the NIMH-funded study was to measure longitudinal change in neuroimaging and neurocognitive measures; however, as a first step we wanted to assess neuroimaging-neurocognitive relationships using the baseline data. We apologize for not including the DTI hypothesis in the introduction, and it is now included:

“We hypothesized white matter disruptions to be associated with worse frontal-executive function and processing speed²³.”

23. Marawi T, Zhukovsky P, Rashidi-Ranjbar N, et al. Brain-cognition associations in older patients with remitted major depressive disorder or mild cognitive impairment: A multivariate analysis of gray and white matter integrity. Biol Psychiatry. 2023;Jun 2(94):12. doi:10.1016/j.biopsych.2023.05.018

- White Matter Hyperintensities: The manuscript references white matter “intensities,” but it is unclear whether FLAIR imaging was included and analyzed. Given the known relevance of white matter hyperintensities (WMHs) in late-life depression, especially in relation to executive dysfunction, it would be good to describe FLAIR acquisition and analysis.

We did not include FLAIR images in our sequence, but in the methods, we clarify that WMH were derived from T1 images:

“In addition, we utilize T1-FreeSurfer-derived total volume of white matter hyperintensities, corrected for total intracranial volume,...

Minor/Specific Comments:

Introduction:

* The initial abbreviation is slightly awkward, consider using "late-life treatment-resistant depression (LLD)".

In Response to Reviewer 3 and Reviewer 2's points, we have revised our reference to LLD as "late life depression" and use treatment resistant LLD when appropriate.

* The statement that "remission may be protective from cognitive impairment" (line 77) is too strong; as shown in this study, the directionality could very well be reversed.

We agree with the Reviewer that cognitive impairment in fact might be a risk factor for non-remission; following the Reviewer's suggestion we have amended this section of the Introduction to talk about 'associations', which is agnostic regarding the directionality:

"Conversely, ~~remission may be protective from cognitive impairment as~~ remitted LLD patients do not show significant differences in brain structure and function from healthy controls, suggesting an association between remission and cognitive remediation."

Results:

* Table 1: Consider reporting non-parametric statistics for skewed variables (e.g., ATHF scores), especially if inclusion criteria imply non-normal distributions.

* Please provide exact p-values for permutation tests where possible.

We have reported exact p values throughout the manuscript unless $p < 0.001$ or when referencing Bonferroni correction (Bonferroni $p < 0.05$) for a correlation matrix.

* Lines 175–178 are long and somewhat repetitive; consider splitting and rephrasing for clarity.

We have broken up this sentence into two based on the Reviewer's suggestions:

"A PLS regression model testing for a relationship between FA in 62 tracts and cognition explained 1.6% of variance in all cognitive tests, significantly more than expected by chance ($p_{\text{PERMUTATION}} = 0.033$, Figure 3A), and FA data specifically predicted Trails A performance but not the performance on any other cognitive test."

Figures:

* Figure 1D: Text suggests this is PLS1, but the figure refers to PLS2—please clarify.

We have updated Figure 1D following reviewer feedback to refer to PLS1; due to a typographical error it was previously referring to PLS2.

* Figure 1G: Not described in the legend—please update.

We thank the Reviewer for highlighting this and have updated the legend to explicitly reference panel G per reviewer suggestion:

"Education showed a small, non-significant association with latent variable brain score (G)."

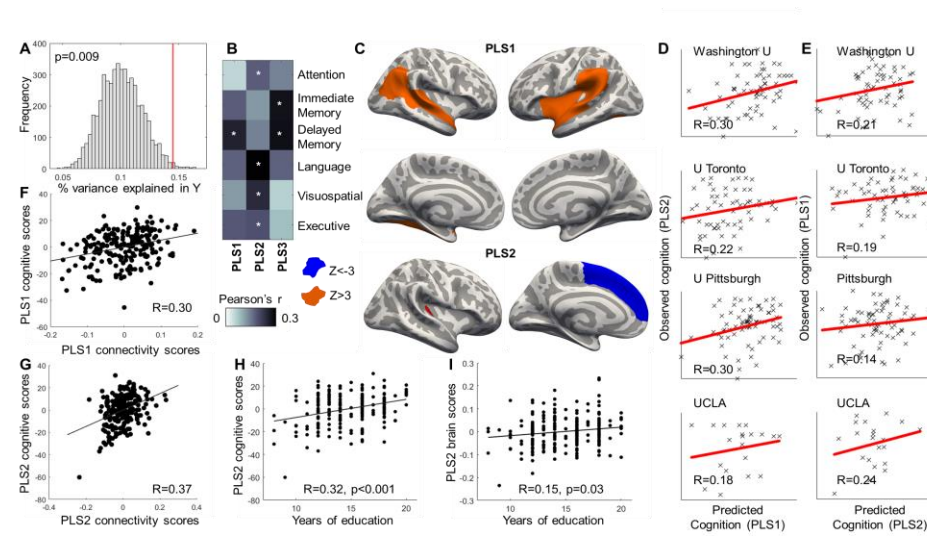
* Figure 2G: Description in the text does not match the content—this needs to be checked.

We thank the Reviewer for highlighting this and have updated Figure 3 legend:

“Education (E) and brain structure centile scores (G) had protective effects on processing speed measured as the PLS latent variable cognitive scores. We did not observe significant associations between education and PLS brain scores (F). “

* Figure 3D: Only one PLS component’s generalisability is shown. Could additional results be added to the supplement?

We thank the Reviewer for bringing up this point. We have now updated Figure 4 to include generalizability data for both PLS1 and PLS2 in the main figure and updated the corresponding figure legend.



”Figure 4. ...”

Supplementary Materials:

* Suppl. Fig. 5: Use vendor names (Siemens, GE) rather than model names (Prisma, etc.), to avoid confusion.

We have amended the vendor names to Siemens and GE to be consistent based on Reviewer feedback.

Reviewer #4 (Remarks to the Author):

In the manuscript "Brain-cognition relationships and treatment outcome in treatment-resistant late-life depression", the authors investigate the relationship between cognitive functioning and neurobiological markers in older adults with treatment-resistant depression. Drawing on data from a multi-center randomized controlled trial, they examine baseline cognitive performance and neuroimaging measures including functional connectivity, white matter integrity, and gray matter structure in a sample of 397 participants, of whom 234 underwent MRI. Using multivariate statistical approaches, including partial least squares (PLS) and elastic net regression, the authors identify brain-cognition associations and attempt to predict treatment remission outcomes. They report that better cognitive function is associated with distinct patterns of brain connectivity and structure, and that poorer cognition and reduced gray matter volume are linked to continued treatment resistance. The study aims to elucidate the neural mechanisms underlying cognitive dysfunction in late-life depression and explore their prognostic value for treatment response.

While the study addresses a clinically meaningful and timely question using a rich dataset and advanced statistical techniques, I have serious doubts about the robustness and interpretability of the reported findings. In particular, the analytical approach appears to rely heavily on in-sample associations without sufficient attention to model validation and generalizability. The use of PLS regression, a technique known to be prone to overfitting when not evaluated in a proper cross-validation framework, is central to many of the conclusions, yet there is no evidence that such validation was appropriately implemented. Moreover, the predictive modeling of treatment outcomes raises concerns about circularity, as it appears to build on features derived from analyses conducted on the full dataset. These issues call into question the stability and reproducibility of the reported associations and predictive effects. As such, I believe the manuscript requires substantial revision before it can be considered for publication.

We thank the Reviewer for their helpful and extensive feedback. We agree with the Reviewer that robustness and generalizability of the models we run is of great importance. In the response below, we would like to assuage some of the concerns raised by the Reviewer by highlighting the robustness analyses that we have implemented in the original submission, and by adding analyses that address the issues raised. Specifically, we highlight the additions we made to methods/results aiming to highlight the following features of the original submission: a) site-based cross-validation of the PLSR analyses b) bootstrapping in PLSR following past work, c) sensitivity analyses used to show consistency of results d) cross-validation of the elastic net logistic regression that we implemented in the original submission. At the same time, we have added new analyses: i.e. a) sensitivity analyses to showcase what is driving some of the brain cognition findings and b) baseline-6m comparisons to showcase the stability over time of the predictors of remission to treatment. We have also further annotated the public github repository.

While we agree that cross-validation analyses are critical, identifying associations within this dataset using robust thresholds and methods is valuable in its own right. Following good practice (Whitaker et al PNAS 2016, Morgan Sarah et al PNAS 2019, Zhukovsky PNAS 2022, and many other papers), we use permutation testing, and bootstrapping with PLS weights >3 or <-3.

In response to the Reviewer's point above regarding the robustness of multivariate models, we agree that multivariate methods like PLS can overfit when the ratio of observations to variables included is low, i.e. too many variables for too few observations in presence of modest effect sizes (Helmer M, Warrington S, Mohammadi-Nejad AR, et al. On the stability of canonical correlation analysis and partial least squares with application to brain-behavior associations. Commun Biol. 2024;7(1).

doi:10.1038/s42003-024-05869-4). However, we ensure that in our models there are more observations than variables, with 210 variables and ~200 participants in rs-fMRI analysis and ~70 variables and ~200 participants in structural analyses. Some literature suggests that strong effects (eg medial temporal volume loss as a correlate of worse memory performance) hold up even in smaller clinical samples (Gratton C, Nelson SM, Gordon EM. Brain-behavior correlations: Two paths toward reliability. *Neuron*. 2022;110(9):1446-1449). We observe this effect in our own work in older adults (Marawi T, Zhukovsky P, Rashidi-Ranjbar N, et al. Brain-cognition associations in older patients with remitted major depressive disorder or mild cognitive impairment: A multivariate analysis of gray and white matter integrity. *Biol Psychiatry*. 2023;Jun 2(94):12. doi:10.1016/j.biopsych.2023.05.018) with robust findings despite the sample N~200.

We also argue that cross-validation approaches in our current predictive models mitigate against circularity. Specifically, we cross-validate each analysis barring the ‘best performing model’. On each cross-validation split fold, we started with the full dataset and train models independently of the test data. The model is then tested on unseen data. The ‘best performing model’ is indeed selecting variables based on the complete dataset, with cross-validation applying to the regression weight testing but not to variable selection. This only affects the ‘best performing model and not the models testing the addition of data blocks (sMRI, fMRI, dMRI, cognition) to clinical data. We now further emphasize this in the Discussion/Limitations as detailed below.

PLS Analysis of Functional Connectivity and Cognition

Latent Variables and Model Specification:

The authors report two significant latent variables (LVs) from the PLS analysis linking functional connectivity to cognition, but it is unclear how the number of latent variables was decided upon. Was this choice guided by prior hypotheses or determined in a data-driven manner? If the latter, what criterion was used?

We have now clarified the LV selection process in the methods:

“Model significance was tested using permutation testing following previous studies^{14,77} (n=5,000). The PLS-R permutation procedure tests whether the model with one, two or three latent variables jointly explains more variance than expected by chance.”

This choice was based on prior studies (Zhukovsky P et al PNAS 2022, Morgan S et al PNAS 2019 and others), whereby we test a 1-LV, 2-LV, and 3-LV model, stopping at 3 LVs given the dropoff in variance explained by each additional LV. In case of the FC-PLS and GM-PLS 3LV solutions were favored, with 3 LVs together explaining more variance than chance. With FA-LS only 1 LV was significant. While sometimes the three LVs jointly explained more variance than chance, we focus on LVs with the most robust weights based on the bootstrapping analysis, presenting those in the main text.

Furthermore, if a second latent variable was extracted, was a deflation strategy applied to remove the variance explained by the first LV? These details are critical for interpreting the structure and independence of the reported LVs.

We agree with the Reviewer, and have added more detail on the algorithm used, which was the SIMPLS algorithm from the *plsregress.m* package with default settings; we added the following to clarify:

Methods:

“We used three partial least squares regressions (*plsregress.m*, SIMPLS algorithm⁷⁷) to identify latent variables capturing multivariate relationships between functional connectivity and cognitive function“

77. S. De Jong. SIMPLS: an alternative approach to partial least squares regression. *Chemom Intell Lab Syst.* 1993;18:251-263.

Generalizability and Overfitting:

The generalizability of the functional connectivity PLS model appears low and is not sufficiently acknowledged in the manuscript. Given that the reported results are likely to be overfitted, particularly in the absence of cross-validation, this limitation should be emphasized more explicitly. In general, PLS results derived without out-of-sample validation tend to be unstable and overly optimistic, and this concern applies equally to the analyses involving white matter and gray matter.

We agree that the generalizability of all the PLS models (when trained on 3 sites and tested on the held-out fourth site) was relatively modest. This result was obtained through a leave-one-site-out cross-validation. We had also tested this with a 5-fold cross validation (random data split, not by site), observing similar levels of cross-validation performance. In the interest of brevity we only report the leave-one-site out analysis. We have amended the results as below:

“Figure 2.... Further, we observed modest good generalizability of the results in hold-out samples (D).”

We have added the generalizability data in Figure 4E for latent variable PLS2 also, quoted in response to Reviewer 3. Finally, we have further qualified our discussion as Reviewer suggested:

Discussion:

“The effect size of brain-cognition relationships identified here was moderate, ~~although relatively strong out-of-sample generalization of the cognitive scores in the FCPLS is encouraging~~ reflected in modest out-of-sample generalization of the predicted cognitive scores.”

Reporting and Statistical Transparency:

There are inconsistencies in how p-values are reported throughout the manuscript. Some are given exactly, while others are expressed as thresholds (e.g., <0.05). I recommend reporting exact p-values in all cases, unless they are below 0.001.

We have updated the intext reporting of pvalues to be the exact pvalue as suggested by the Reviewer and other Reviewers, unless they were below 0.001. We already report exact p values in the Figures.

Additionally, the manuscript does not sufficiently describe the permutation test procedure used to assess the significance of the latent variables. What exactly was permuted in these tests? What were the p-values for LV1 and LV2? The method and rationale for associating individual functional connectivities or cognitive tests with the latent variables are also not clearly described. Was a statistical significance test applied to these associations, and if so, were multiple comparisons corrected for?

We thank the Reviewer for bringing this up. To the first part of the question, we have added more information on the permutation procedure:

Methods:

“Permutation involves random row-wise reshuffling of the Y data on each permutation, and then running a PLS regression, while keeping the correlation structure of the original X and permuted Y data.”

On the second part of the question (p-values for each LV), the significance testing is for the 1, 2 or three components jointly as outlined in response to the Reviewer’s previous point. As such it’s a limitation of this approach; the overall/joint model significance might be driven by one LV, making all

three jointly significant but this being mostly due to eg LV3. This is inherent to the PLS model identifying a pattern using all latent factors included and is common throughout the prior PLS literature, at least with this matlab implementation.

On the third part of the question, we did apply Bonferroni correction on the correlations between XS latent variable scores and Y. We state so currently in each figure legend but also now added this to the methods.

Figure legend:

“Univariate Pearson’s correlations between individual cognitive tests and latent variable scores are shown in (B, * $P_{\text{BONFERRONI}} < 0.05$)”

Methods

“We correlated the latent brain scores (XS) with the cognitive tests and applied Bonferroni correction ($P < 0.05$) to identify cognitive tests significantly associated with each latent variable.” (underlined added)

Discrepancies and Missing Correlations:

There is a notable inconsistency between the reported explained variance of 8.4 percent and the explained variance shown in Figure 2A. Moreover, the authors do not report the correlations between PLS1 Brain and PLS1 Cognition, or between PLS2 Brain and PLS2 Cognition. These values are fundamental to interpreting the strength and relevance of the latent variables. It is also unclear why the manuscript focuses so heavily on the first PLS mode while omitting a parallel discussion of the second.

We thank the reviewer for pointing this out this issue. The first iteration of these models we ran included three PLS LVs, which is what Figure 1A is based on. We had to exclude the Columbia site from the analyses as noted in the methods, and we reran the models, resulting in two LVs being jointly significant. The three LV model had a pvalue of 0.07. The results with and without Columbia were highly consistent as it was only one of five sites, and the LV shown in Figure 1 was also consistent across the 2 and 3 LV models. We apologize about missing the %variance histogram in Figure 1A and have updated it now.

We found PLS1 to be more strongly correlated with the cognitive data Y, hence we focus on PLS1 and not PLS2. In fact, PLS1 explains 6.3% of variance in Y, while PLS2 explains only 2.06%. We appreciate the utility of showing the correlation between both PLS1 and PLS2 and Y, so the reader has a full picture of the association pattern being driven by PLS2 and have added this to Figure 1B following Reviewer’s suggestion. We thank the Reviewer for bringing this up as we believe it makes for a more comprehensive report of our FC-PLS findings:

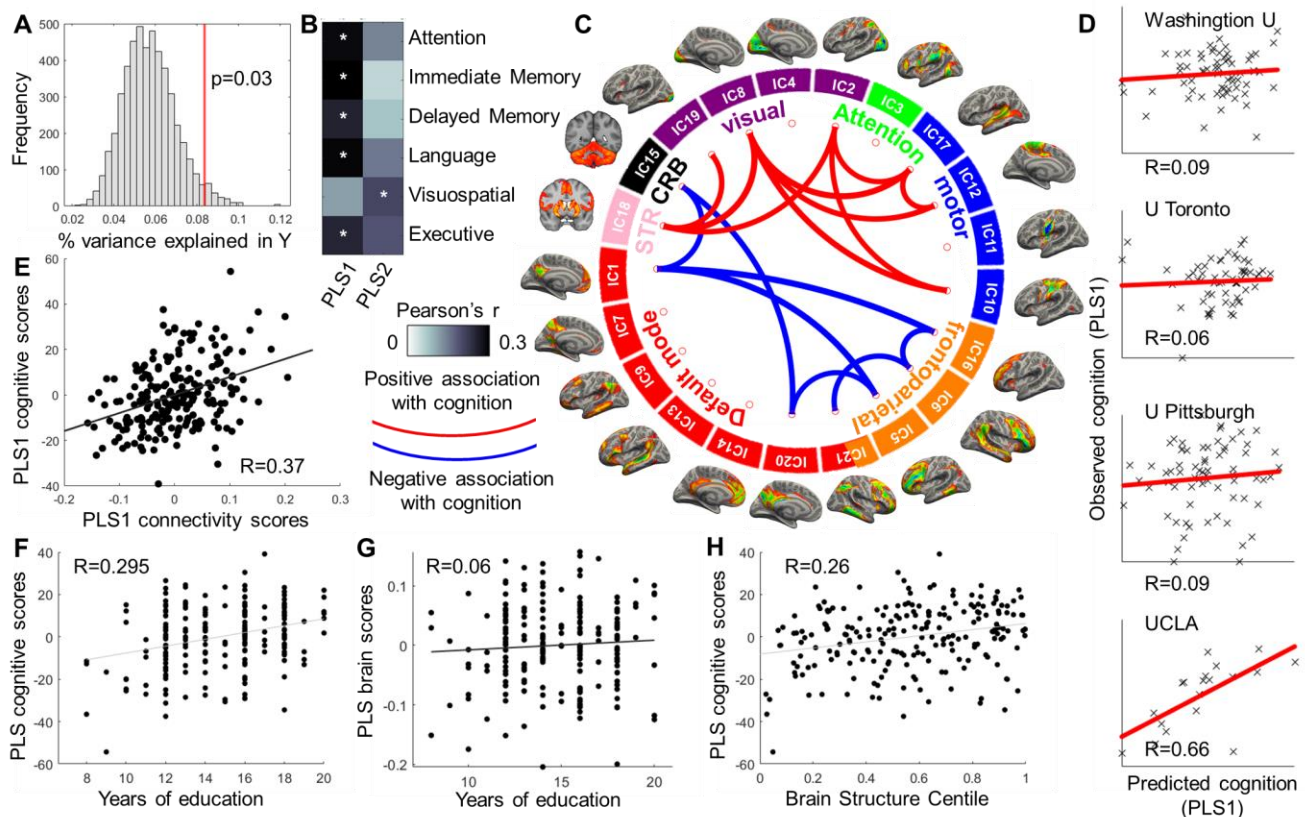


Figure 2....

Figure Presentation and Scatter Plots:

Figures 2D, 3D, and 4D contain scatter plots that are difficult to interpret and appear inconsistent across modalities. It is unclear what is being shown in each case. For example, the rationale for plotting the correlation between observed and predicted cognition is not explained, and it is not clear how the "observed cognition" scores were calculated. Typically, PLS results are illustrated by plotting the brain and cognition latent scores against one another to demonstrate the strength of the mode. The authors do not show these brain-cognition score correlations, which would be the appropriate way to assess model quality and generalizability. Additionally, all scatter plots shown appear to reflect in-sample fits. There is no cross-validated version of these plots or associated metrics, which further limits confidence in the robustness of the findings.

We thank the Reviewer for bringing this up – Figures 2D, 3D and 4D are indeed cross-validated, with training in 3 sites and testing in the fourth site, completely unseen to or held-out from the training data. We now clarify what cognitive scores are shown in each figure in the figure legend.

Figure 2D legend:

“... Further, we observed modest generalizability of the results in hold-out samples (D). Models data were trained on data from three of the four sites. We indicate held-out sites used for model testing in (D). Predicted cognitive scores were calculated as the first principal component of variation in the five cognitive tests significantly associated with PLS1 in (B) to ensure consistency across cognitive scores derived from all data splits....”

Figure 3 legend

"We observed good generalizability of the results for Trails A performance in all held-out samples (C)."

Figure 4 legend

"We observed good generalizability of the prediction of in most hold-out samples (D, E). Specifically, leave-one-site-out cross-validation analyses were performed to predict the first principal component of variance in the cognitive tests significantly associated with brain structure in (B)."

Predicting Treatment Outcomes

The section on predicting treatment remission raises several important concerns regarding the validity and transparency of the modeling approach. While the authors state that they tested their best-performing model in an out-of-sample fashion, only a single AUC value is reported, and it is unclear how this was derived. Although the methods section describes the use of cross-validation, the results do not reflect a proper cross-validation framework. There is no indication that performance was averaged across folds, nor are confidence intervals or variability measures reported. This lack of transparency makes it difficult to assess whether the results are truly generalizable.

We tried to be explicit about the variable selection process for the best performing or most parsimonious model including variables that were selected across the whole sample but have now added more detail as requested. The remaining discovery prediction models were all cross-validated with variable selection not experiencing any leakage between train and test data.

"Three sets of prediction models were run, each including demographic, clinical, cognitive data and one of the three imaging modalities. The clinical model included age, sex, and baseline MADRS scores. These models were run in the larger sample (n=397) including cognitive data, clinical data, and demographic data. Then the model was re-run in the sample (n=234) that also included neuroimaging to determine the potential improvement in the AUC when adding neuroimaging. Stringent cross-validation approach ensured that model training was completely separate from the test data. Following this iterative process, we selected the most parsimonious model that included predictors surviving regularization in over 95% of models ran. We then tested the performance of this most parsimonious model in an even 8-fold and 5-fold split of the step 1 and step 2 data, respectively, presenting confusion matrices of this model showcasing the true positives and negatives as well as false positives and negatives. Feature selection in this most parsimonious model was affected by train and test data, with cross-validation applying to the regression weights but not to the variable selection as variable selection process was done across the whole sample."

In fairness, we believe the true predictive performance likely lies in between the fully cross-validated feature selection (auc=0.74) and the best performing model performance (auc>0.8). The discovery models search through all 74 MRI cortical thickness and volume variables, many of which we have no apriori constrained hypotheses and so those (eg visual cortex thickness) could be discarded similar to the most parsimonious model.

Moreover, it is not clear whether features derived from the PLS analyses were included as predictors in the outcome model. If so, this would raise serious concerns about circularity or "double-dipping," unless the entire PLS procedure, including latent variable estimation and selection, was embedded within the cross-validation loop. As currently reported, there is no evidence that this was the case. Without such safeguards, the reported predictive performance is likely inflated and may not hold in independent samples. The authors need to clarify this modeling pipeline in full detail and, if necessary, re-run the

analysis using a fully nested cross-validation framework that includes all preprocessing and dimensionality reduction steps.

We appreciate the Reviewer's emphasis on cross-validation and share this passion for robust models. In our original analyses, the PLSR model predicting change in depressive symptoms included all variables, 74 structural MRI, three cognitive variables and clinical/demographic variables. We show cross-validation performance in Figure 5 (completely unseen test data, not pre-selecting variables, no leakage between train and test data: $r=0.35$ in Fig 5H, considerably lower than within sample $r=0.63$, but still relatively high).

Figure 5 legend:

"A partial least squares (PLS) regression including demographic, clinical, cognitive and all available brain structure predictors explained a significant amount of variance (39.7%) in slopes of change in MADRS scores (G), with moderate predictive accuracy in held-out data (H)."

General Conclusion

In light of the concerns outlined above, particularly regarding the lack of appropriate cross-validation, potential circularity in predictive modeling, and limited transparency in statistical procedures, I recommend that the manuscript undergo a major revision. The authors should re-implement their multivariate analyses and predictive models within a rigorous cross-validation framework that includes all preprocessing and dimensionality reduction steps. Permutation testing should be applied not to in-sample associations, but to the cross-validated model performance, in order to assess the statistical significance of generalizable patterns (<https://doi.org/10.1016/j.bpsc.2022.07.012>, <https://doi.org/10.1016/j.neubiorev.2020.10.030>). Addressing these issues would significantly strengthen the validity and interpretability of the findings and improve the manuscript's contribution to the field.

We appreciate the Reviewer's emphasis on permutation testing of the cross-validated data. We argue that our current models follow best practices in the field for showing in-sample significance, with robust thresholds and multiple comparison correction. Given the extensive analyses performed and cross-validation correlations shown, we believe this includes sufficient evidence of generalizability (modest for FC and FA but good for cortical thickness, and high for the step 1 treatment outcome prediction, modest for step 2 outcome/remission prediction). We have referenced the Mihalik et al article above in the methods to emphasize out of sample prediction importance:

"We applied the PLS beta regression coefficients obtained in the training sample to the test sample and correlated the observed cognitive data with the predicted cognitive data to assess PLS performance in predicting cognitive function in held-out data⁷⁹."

79. Mihalik A, Chapman J, Adams RA, et al. Canonical Correlation Analysis and Partial Least Squares for Identifying Brain–Behavior Associations: A Tutorial and a Comparative Study. *Biol Psychiatry Cogn Neurosci Neuroimaging*. 2022;7(11):1055-1067. doi:10.1016/J.BPSC.2022.07.012

We were not clear how the other reference provided by the reviewer was relevant for PLS method so we did not cite it:

"Neural mechanisms of visual sensitive periods in humans"
<https://doi.org/10.1016/j.neubiorev.2020.10.030>

Remarks on code availability:

The code repository associated with this manuscript is unfortunately quite difficult to navigate and

currently lacks sufficient documentation. Given the complexity of the analyses and the use of multiple neuroimaging modalities and statistical frameworks, a clearly structured and well-documented codebase is essential to ensure transparency and reproducibility. At present, it is challenging to trace how key results were generated or to verify the analytical pipeline as described in the manuscript. I strongly encourage the authors to substantially reorganize the repository, adopting a more intuitive folder structure and providing a comprehensive README file. This should include clear, step-by-step instructions on how to reproduce all major analyses and figures from the raw or preprocessed data. Improving the clarity and usability of the codebase will significantly enhance the value and credibility of the study.

We thank the Reviewer for highlighting this and have added more documentation to the codebase. We refrain from including a notebook style documentation but have added more notes to each of the steps in the code.

Thank you for the positive reply to our resubmission of the manuscript “Brain-cognition relationships and treatment outcome in treatment-resistant late-life depression”. We have addressed the outstanding issues raised by reviewers in the comments below and substantially improved our remission prediction figure, rerunning analyses and re-plotting the results (including distributions), in addition addressing other queries from the editors such as publishing code as a code release on Zenodo. We hope that the manuscript is now in compliance with the Nature Communications guidelines,

We look forward to your response,

Aristotle N Voineskos, MD, PhD (on behalf of all authors)

REVIEWERS' COMMENTS

Reviewer #1 (Remarks to the Author):

All my concerns have been addressed.

We thank the Reviewer for their efforts to help us improve this manuscript across the review period.

Reviewer #2 (Remarks to the Author):

I only have one remaining recommendation, pertaining to the Results (comment #4). As requested, the authors reanalyzed the brain-cognition associations within each subgroup (cognitively unimpaired, MCI + dementia). For all models (rsFC, diffusion-weighted, and T1) the overall PLSR was non-significant in the subgroups (in contrast to the findings when the subgroups are combined). As the authors note, this suggests that the brain-cognition findings are driven by variability across the cognitive subgroups, and thus the models require the variability provided by including patients with various levels of impairment, and may not perform similarly if restricted to more homogenous groups alone (e.g., cognitively unimpaired alone). This is a critically important finding when considering generalizability and for implementation in future work. While these additional findings (and interpretation) were added to the Supplement, they should also be integrated into the primary manuscript to ensure these important results are emphasized to readers

We thank the Reviewer for highlighting this issue. In response to Reviewer 3 we also re-ran the PLS regressions showing that regressing out cognitive status from cognitive outcome data Y altered model significance. We have now included an explicit reference to this finding and its implications in the discussion as per the Reviewer's helpful suggestion (quoted text below; added text underlined):

Discussion

“The effect size of brain-cognition relationships identified here was moderate, reflected in modest out-of-sample generalization of the predicted cognitive scores. Sensitivity analyses showed that our brain-cognition findings were only significant in the combined group, suggesting that greater heterogeneity in cognitive performance⁶² across cognitive subgroups helps identify brain-cognition associations.”

62. Gratton C, Nelson SM, Gordon EM. Brain-behavior correlations: Two paths toward reliability. *Neuron*. 2022;110(9):1446-1449.

Reviewer #2 (Remarks on code availability):

N/A

Reviewer #3 (Remarks to the Author):

I thank the authors for their revised work and answers to my questions. However, I still have some concerns regarding the timing of data acquisition (in line with the questions of reviewer 2 as well).

The argument that post-treatment MRI and cognitive measures can be treated as baseline predictors because baseline and 6-month values correlate strongly is not entirely convincing. High correlations primarily indicate rank-order stability across participants and do not rule out systematic treatment-related changes. Moreover, the statement that treatment would affect MRI and cognition in the same way is not well supported by the literature. Structural brain measures and cognitive outcomes may change on different timescales or show different sensitivities to treatment. Therefore, the justification for including measurements obtained after treatment onset as predictors of treatment response remains insufficient.

We acknowledge the Reviewer's point that treatment may not impact MRI measures and cognition measures in the same way, and that these effects may occur at different timescales. At the same time, we are aware that studies of Alzheimer's and MCI progression do suggest a longer timescale (in the order of years) for both cognitive and structural gray matter changes in the course of MCI progression, with most impactful studies looking at 10+ year long follow-up time periods. The cohort we report on here is part of a larger longitudinal study that also includes 6 month and 24 month MRI and cognitive outcomes. These longitudinal data are currently being analyzed, although we have now included preliminary analyses of some of the 6 month data in response to prior reviews. Our discussion section currently states this limitation, ensuring caution regarding the DTI and rs-fMRI remission prediction models, and draws attention to the opportunity for further analyses in future work that can further assess issues regarding timescale and sensitivities to treatment (deleted text and added text highlighted and underlined):

"Further, although cognitive function and brain MRI were assessed at varying time intervals relative to the OPTIMUM treatment, our supplementary data show that brain structure and cognition were remarkably stable over a 6-month time period. It is possible that OPTIMUM treatment has impacted functional and diffusion MRI, and we found these markers did not predict remission. Future studies using a more controlled design may therefore identify connectivity markers of remission in LLD. Both brain and cognitive data were collected at the same interval relative to the OPTIMUM treatment, supporting our cross-sectional brain cognition findings. Future work analyzing longitudinal data collected at later time points will further illuminate change in MRI and cognitive measures over time as well as prediction of treatment

[response over time. Such work can help further clarify issues of potentially different timescale or sensitivity to treatment of structural brain measures and cognitive outcomes.](#)

The additional subgroup analyses separating participants with and without MCI are appreciated. However, rerunning the PLS analyses in substantially smaller samples mainly results in a loss of statistical power, making it difficult to draw conclusions from the non-significant findings. The concern is more related to whether the brain–cognition relationships differ as a function of cognitive status. This question might be addressed more directly by incorporating MCI status as a grouping variable or covariate within the PLS framework, rather than splitting the sample. Such an approach would allow the authors to test whether the patterns differ between groups while retaining the full sample size.

We thank the Reviewer for this suggestion and agree that it's an excellent way to disentangle the effects of cognitive group vs the reduction in power from splitting the sample into two similarly sized groups. We re-ran the PLS after correcting for HC vs MCI groups in the cognitive outcome (Y) data and found that while within-HC and within-MCI correlations are largely similar after correcting for group, the overall across-group correlation is reduced from 0.37 to 0.33, resulting in a p value of $p=0.03$ changing to $p=0.20$. This is consistent with our subgroup sensitivity analysis. We have now added a figure and more detail to the supplement and in the discussion:

Discussion (added text underlined):

““The effect size of brain-cognition relationships identified here was moderate, reflected in modest out-of-sample generalization of the predicted cognitive scores. Sensitivity analyses showed that our brain-cognition findings were only significant in the combined group, suggesting that greater heterogeneity in cognitive performance⁶² across cognitive subgroups helps identify brain-cognition associations.”

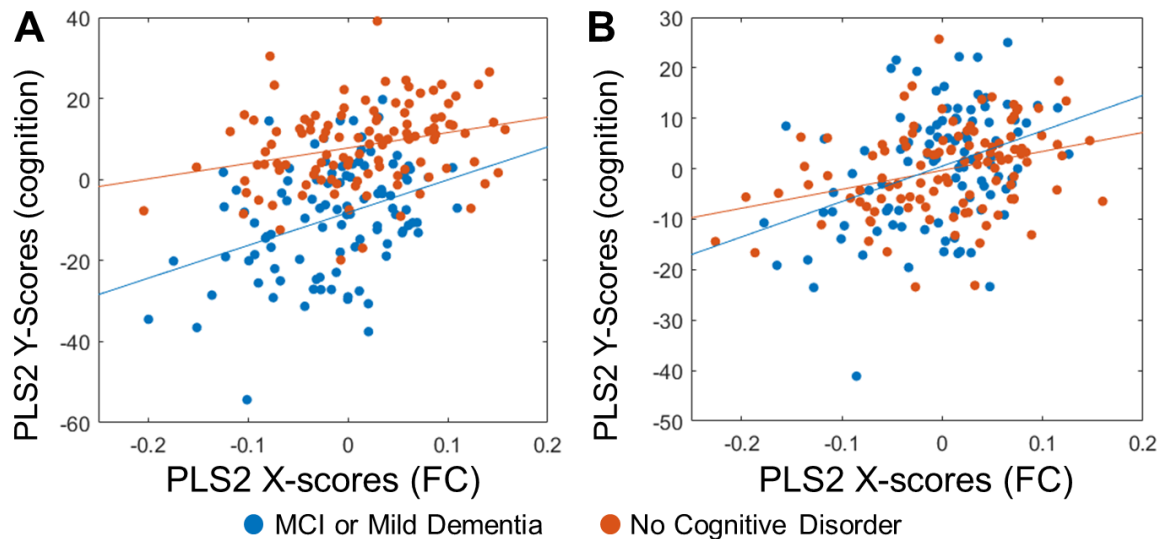
Supplementary information (added text underlined)

“10. Brain-cognition associations in subsets of participants and after controlling for cognitive status

We next tested whether our findings in the overall OPTIMUM sample hold up in subsets of participants with MCI or mild dementia, and in those with no cognitive disorder. For the resting-state FC models, we found that in each subgroup individually, the overall PLSR was not significant, i.e. did not explain more variance in cognitive function compared to chance permutation models (cognitively normal subjects: $n=112$, $p_{\text{PERMUTATION}}=0.74$ for a 3-latent variable model; MCI and dementia subgroup: $n=97$, $p_{\text{PERMUTATION}}=0.36$ for a 3-latent variable model). For the diffusion-weighted fractional anisotropy models, we found that in each subgroup individually, the overall PLSR was not significant (cognitively normal subjects: $n=109$, $p_{\text{PERMUTATION}}=0.30$ for a 1-latent variable model; MCI and dementia subgroup: $n=103$, $p_{\text{PERMUTATION}}=0.12$ for a 1-latent variable model). For the T1 cortical thickness and subcortical volume models, we found that in each subgroup individually, the overall PLSR was not significant (cognitively normal subjects: $n=113$, $p_{\text{PERMUTATION}}=0.28$ for a 3-latent variable model; MCI and dementia subgroup: $n=103$, $p_{\text{PERMUTATION}}=0.54$ for a 3-latent variable model). This suggests that our brain-cognition findings could be driven by the variability in cognitive function across the two patient subgroups and reduced power to detect effects in the smaller cognitive

subgroups. Greater variability in brain data and cognitive/behavioral performance drives larger effect sizes⁴ and leads to significant findings in the whole sample.

Finally, we repeated the brain-cognition PLS regressions after controlling for the “no cognitive disorder” vs MCI/Mild dementia group in the brain connectivity data using ComBat and after regressing out cognitive status from the cognitive outcome variables. The resting-state FC PLS remained significant (permutation $p=0.0286$), similar to the DTI PLS (permutation $p=0.0386$) and the CT PLS (permutation $p=0.0138$) after controlling for cognitive status in the brain data. However, regressing out group from the cognitive outcomes Y did affect the models as shown in Supplementary Figure 9. Specifically, we found the variability in cognitive outcomes Y to be reduced after regressing out cognitive status, leading to lower overall PLSR correlation of 0.33 vs 0.37.



Supplementary Figure 9. Latent variable correlation in the NCD (no cognitive disorder) and the MCI or Mild Dementia groups for the FC PLS run across all participants before **(A)** and after **(B)** regressing out group from the cognitive outcome data. We found a significant correlation between latent brain and cognitive scores from the resting-state FC PLS regression when not regressing out group **(A)**, Pearson’s $R=0.2748$, $p=0.0034$ in NCD, Pearson’s $R=0.3675$, $p=1.8e-04$ in MCI/Dementia, overall $R=0.3704$). After regressing out group, we still observe significant correlations within cognitive subgroups, but the overall PLS model is no longer significant (Pearson’s $R=0.3215$, $p=5.5e-04$ in NCD, and Pearson’s $R=0.3689$, $p=1.7e-04$ in MCI/Dementia, overall $R=0.3334$, permutation $P=0.2038$ for the overall PLS regression).”