We would like to thank the reviewers for their valuable input and overall appreciation of our work. We have revised the manuscript accordingly and highlighted the changes in yellow. Please find below a point-to-point reply to the reviewers’ comments.

**Reviewer 1 Comments for the Author...**

In this really interesting study authors are using innovative AI algorithms to estimate chronological age of different group subjects from their neuroimaging data. Interestingly, they are comparing MRI-BAG and FDG PET-BAG as imaging biomarkers of neurodegeneration. Moreover, they explore the associations of these two parameters with cognitive performance and Alzheimer pathology (amyloid and tau), and an estimation of the probability of cognitive outcome value by means of these parameters.

1. **Unfortunately, the study contains major methodological flaws related to the statistical analysis. Authors extracted an equally sized subsample of stable subjects matched by age and sex to the cohort of all decliners (both in CN+SCDADNI and MCIADNI subjects). Instead of conventional logistic regression, this approach needs a conditional logistic regression analysis due to the presence of matching. Moreover, a case control study like it is proposed here (similar number of subjects matched by age and sex) allows to estimate the Odds Ratio but not the probability of cognitive outcome. For this purpose, a study of cohorts (using the whole population available) would allow to estimate the real probability of conversion. Therefore, it should be discussed with their statistician if this study have to include the whole population adjusted by age and sex as covariates and use a conventional logistic regression analysis to model the relationship between BAG from the significant imaging modality and cognitive outcome.**

The reviewer has made the important point that true a priori probabilities for conversion in the whole population are not equal to 50%, which was implicitly assumed by creating a case-control study. After thorough discussion, we have revised our logistic regression analyses and made the following three changes:

1) As suggested by the reviewer, we now included all participants for whom information about conversion was available in the prediction model.

2) Additionally, we investigated and present model calibration in the revised version of the manuscript to XXX.

3) Even the frequency of conversion observed in the training data does not necessarily represent a priori probabilities well and thus biases the estimation of posterior probabilities. To counter this pitfall, we have calibrated the posterior probabilities estimated from the logistic regression model with the iterative expectation maximization as suggested in [1].

1. **Some of the statistical methods are described in the specific sections of comparisons and correlations performed. However, there is a lack of section where all the statistical analysis should be detailed. Moreover, the description of r of Pearson and rho of Spearman analysis only appears at the results section.**

We have combined the description of all statistical analyses to a unique section (p. 7), including justification for the usage of the Pearson or Spearman method for correlation.

1. **To derive a BAG cutoff, authors intercept of the logistic regression curve at 50% probability. However, the use of this cutoff might be arbitrary. Instead, you should use AUC ROC graphs for optimal cutoff of maximal sensitivity and specificity.**

Following this comment, we changed from using 50% probability as threshold to the more dynamic option of observing AUC ROC, and, importantly, also precision-recall curves to infer an optimal cut-off for pending cognitive decline.

1. **Please present the mean difference and confidence interval instead of t statistic in the respective paragraphs of the result section.**

We now present the mean difference and confidence interval instead of the t-statistic.

1. **PPV and NPV can´t be estimated in case control studies unless you used specific methods (i.e. Mercaldo).**

Following comment #1 by reviewer 1, we refrained from performing a case control study (see answer to comment #1).

1. **Please include confidence intervals for all indexes described.**

We now present confidence intervals for all indexes described.

1. **The term marginal is frequently misleading in the text. It should be used “very weak but almost significant”. Take into account that very big sample sizes lead to a lot of irrelevant but statistically not significant findings.**

As described in the new “statistical analyses” section (p. XX), we now consider the following significance levels: p < .1 = “almost significant”, p < .05 = “significant”, p < Bonferroni correction = “significant after Bonferroni correction”. Due to the multitude of analyses conducted and JNM’s word limitation, we did not incorporate the whole term “very weak but almost significant”.

1. **Use “FDG PET BAG” instead of “FDG PET” when referring specifically to this particular parameter, (FDG PET is a more general term which offers some additional important parameters). Example: “… both MRI- and FDG-PET-BAG significantly predicted cognitive outcome in MCIADNI, while only FDG-PET predicted cognitive outcome in CN+SCDADNI (see Tables SM2-SM5 for estimates of logistic regression in sample 1 analyses using only one imaging modality).”**

We have changed all instances of “FDG PET”, where indeed “FDG PET BAG” was referred to, to “FDG PET BAG”.

1. **Please use “and” or “or“ instead of a “slash” in the text. These are two examples:**

**• “… cognitively normal (CN)/had subjective cognitive decline (SCD)”. You can use CN/SCD as an acronymous but after an explanatory text without symbols.**

**• “… with cognitive performance/AD neuropathology in these cohorts.”**

In the revised version of the manuscript, we have disentangled the groups CN and SCD, where possible (see answer to comment #14). All other positions in the text that previously contained a slash were corrected as per the reviewer’s suggestion.

1. **Please explain why you didn’t consider to add the Ratio CSF AB-pTau to the single AB or pTau in CSF.**

Thank you for this interesting remark. Research has shown that CSF p-Tau/Aβ1-42 ratio is strongly correlated with PET amyloid load [2], which is the gold standard for *in vivo* amyloid assessment. This correlation indeed outperformed single measurements of t-Tau, p-Tau and Aβ1-42 in CSF. Following the reviewer’s suggestion, we have thus added CSF p-Tau/Aβ1-42 ratio to the variables of interest in the correlation analyses (p. XX). For the prediction of cognitive outcome, we have XX replaced XX Aβ1-42 with CSF p-Tau/Aβ1-42 ratio so the expected correlation between Aβ1-42 and CSF p-Tau/Aβ1-42 ratio would not interfere with the variables’ explanatory power for cognitive outcome [3] (p. XX).

1. **Some discussion about the specific cortical areas that have been shown to be prognostic in other studies (ie. Posterior Cingulate, Parietal for FDG-PET, and hippocampus for MRI) are missed. Besides, a comparison with other papers which explored the brain hypometabolism in FDG PET and cortical atrophy in MRI should be convenient.**

XX To do…

**Reviewer 2 Comments for the Author...**

Doering et al. applied machine learning algorithms to structural MRI and FDG-PET images of n=376 elderly subjects with subjective cognitive decline (SCD) and without to calculate brain age gap (BAG). The authors associated MRI- and PET-derived BAG with some cognitive tests, biomarkers of Alzheimer’s disease (AD), and with a cognitive outcome in 2 years. They report correlations between BAG and amyloid-beta in cerebrospinal fluid in the above subjects and in patients with mild cognitive impairment (MCI), as well as between BAG and cognitive performance in MCI patients. Furthermore, PET-derived BAG predicted cognitive deterioration in SCD+healthy subjects, while MRI-derived BAG predicted cognitive deterioration (to dementia) in MCI patients. Doering et al. conclude that BAG can be reliably estimated from FDG-PET and MRI images. Whereas PET-derived BAG is more sensitive to cognitive deterioration in subjects without objective cognitive impairment, MRI-derived BAG is indicative of impending dementia in patients with MCI.

Major criticism

1. **Application of the results. As the authors themselves note, BAG has been defined so far using MRI data. This is understandable, since FDG-PET is not indicated in cognitively healthy subjects. Hence, application of PET-derived BAG is basically limited to clinical trials (and academic studies). Specifically, the authors note that “FDG-PET BAG could complement the identification of at risk individuals, as individuals with a BAG below our proposed cuttoff are unlikely to develop cognitive impairment within two years”. This statement is based on findings in an external cohort of SCD subjects (“DELCODE”), where sens, spec, PPV and NPV of 88%, 34%, 13%, and 96% are reported. Yet, this cohort include n=80 cognitively stable and only n=8 cognitively deteriorated subjects. Thus, there is per definition a bias toward high NPV. Remarkably, in an equally weighted (n=30 stable vs. n=30 decliners) subset of the initial cohort, i.e., the cohort that was used to derive PET-BAG, sens, spec, PPV, and NPV were only 70%, 67%, 68%, and 69%. I’m afraid, these results are not sufficient to recommend PET-derived BAG for the use in clinical trials. Consistently with this limited predictive power of PET-derived BAG, it either did not correlate or correlated only marginally (r=-0.100, p=0.06) with the cognitive tests.**

We agree with the reviewer that our results are not sufficient for the stand-alone usage in clinical trials. In our work, we intended to investigate the value of PET for the estimation of brain age, as well as the clinical utility of PET-estimated BAG. Thus, our work provides a step forward in the direction of establishing clinically useful biomarkers based on brain age, rather than being the whole story. The lack of strong correlations of BAG and cognitive tests is to be expected in a cognitively unimpaired cohort, as fluctuations in these scores within the realm of a “cognitively normal” diagnosis have proven not to be meaningful. Yet, we have phrased our discussion pertaining to the utility of FDG-PET BAG more carefully as suggestions/stepping stones, rather than stand-alone recommendations on p. XX:

XX

Furthermore, we have adapted our title to reflect this alteration to XX.

1. **Irrespective of the above, what is the point of using PET-derived BAG instead of PET itself as marker of cognitive decline? See e.g., Scheef et al., 2012. The former requires rather complex analyses including machine learning, while analytical pipelines of FDG-PET data are well established. In the same vein, MRI features alone might predict cognitive decline in SCD subjects, see Ebenau et al., 2022. To summarize the first two comments, I question the utility of BAG in general and PET-derived BAG in particular as marker of cognitive decline in clinical trials. To justify this application, the authors should compare BAG with established regional features of FDG-PET, MRI images (e.g., hippocampal volume), and with chronological age as reference. This is equally true for MCI patients. Otherwise, another application of (in particular PET-derived) BAG should be proposed.**

Our analyses were not solely intended to investigate MRI or FDG-PET BAG for cognitive decline, but rather to inspect the utility of MRI or FDG-PET BAG in reflecting normative brain health at a particular age as measured in cognitive dysfunction and pathology. We hypothesized that cognitive decline could possibly be considered a consequence of a less healthy brain, i.e., accelerated brain aging, which is why we also used BAG to predict cognitive outcome. We were able to show that XX repeat findings XX, while using the two modalities in separation, thus showing that MRI and FDG-PET BAG can serve as independent normative markers of a healthy brain, which in turn could be an additional, rather than stand-alone, marker to be assessed in clinical trials, where these modalities are already being collected. To investigate the complimentary value of BAG in the prediction of cognitive outcome, we split our logistic regression models in two parts: one with only FDG-PET features, where we added XX Scheef features XX to the prediction of cognitive outcome using FDG-PET BAG and one with only MRI features, where we added XX Ebenau features XX to the prediction of cognitive outcome using MRI BAG.

1. **The main study cohort represents a mixture of SCD subjects and cognitively healthy subjects without cognitive complaints. As the authors themselves note, SCD subjects are more likely to develop MCI and dementia due to AD, for a recent meta-analysis see Pike et al., 2022. Thus, BAG is biased towards disease-related acceleration. The authors should treat these groups separately or exclude subjects with SCD. Inclusion criteria should be clearly stated. In particular, what is the status of cognitively healthy subjects regarding the AD biomarkers?**

It is a highly relevant question of whether or not CN and SCD should be grouped together in our analyses. While previously, we kept CN and SCD grouped together to increase both sample size and variance of our training sample, as well as due to methodological flaws in ADNI (see below), we now agree that it is important for our research question to treat CN and SCD as separate groups. In the revised version of the manuscript, we trained the brain age estimation models only on data from CN (n = 276) and reserved SCD (n = 102) from ADNI as an additional clinical sample. Note, however, that the label “SCD” was only introduced in ADNI2. Individuals recruited during ADNI1 (~1/4 of our sample) may therefore have had SCD which was undetected at the time and which possibly caused the indifferent results of brain age between CN and SCD in the ADNI sample. Exclusion of these individuals would cause further shrinkage of our sample size, which would be undesirable, especially pertaining to analyses of cognitive decline.

Based on

1) the above observation that CN and SCD cannot be completely disentangled in ADNI,

2) the fact that MAE, mean BAG and the frequency of conversion were highly similar across ADNI CN and SCD (see below table)

And 3) the fact that the group of cognitive decliners among just the SCD group (n=10 decliners from 83 subjects where longitudinal data was available) was too small to predict cognitive outcome in this cohort (see deviation of points from “perfect calibration” in below calibration plot (figure a)),

we grouped CN and SCD together to a cognitively unimpaired group (CU) for analyses of correlation between BAG and cognitive performance, AD neuropathology and the prediction of cognitive outcome. The calibration plot of this sample is presented in figure b) and shows obvious improvement over a).

To summarize, while our models were trained on individuals who were cognitively normal and without any report of subjective impairment, for subsequent analyses, the small sample size and methodological pitfalls of early ADNI phases with regards to SCD characterization required to group CN and SCD to CU.

We added the following to the limitations section to acknowledge the lack of data on SCD for better within-group analyses (p. XX):

*“Moreover, there is a lack of publicly available big neuroimaging databases on SCD, enabling to disentangle early differences in brain health, possibly related to cognitive decline, between CN and SCD. The SCD label was only included in the second phase of the ADNI study – individuals recruited during ADNI-1 (~1/4 of our sample) may therefore have had SCD which was undetected at the time and which possibly caused the indifferent results of brain age between CN and SCD in the ADNI sample. However, exclusion of these individuals would have caused further shrinkage of our sample size, which would have been undesirable.”*

**Further comments**

1. **It is not plausible that MRI- rather than PET-derived BAG predicted cognitive decline in MCI patients. Numerous studies reported FDG-PET to predict cognitive decline in a more sensitive manner than MRI. The information on chronological age is the same in both BAGs. How do the authors explain this finding?**

As discussed in the answer to comment #13, our models were not (intended to be) trained to predict cognitive decline, but rather, to provide a normative marker of brain health. We have shown that BAG estimated from FDG-PET and MRI, and therefore, age-relevant brain regions in FDG-PET and MRI show different sensitivities to cognitive decline. We thank the reviewer for pointing out that this information was insufficiently discussed in the previous version of the manuscript and have stated it more clearly on p. XX:

XX

1. **Why only 60+ subjects were included? Accelerated aging to be captured by BAG should begin earlier.**

While it is true that accelerated aging, potentially indicative of future cognitive impairment, may begin before the age of 60, the decision to only consider individuals older than 60 was made based was based on the following two reasons:

1) A data-driven reason to exclude subjects younger than 60 was that few subjects exist in ADNI at this age range, thus model reliability for this age range would be compromised, and the accuracy of the whole model could suffer from their inclusion.

2) In [4], Jessen et al. provide recommendations for the characterization of SCD. They state that “In individuals younger than 60 years of age, the likelihood of a medical condition causing future cognitive decline and dementia is low, which suggests that the likelihood of SCD in individuals younger than 60 years being related to other or potentially reversible causes (eg, depression) is higher than in individuals aged 60 years or older.“ To decrease variance of potential underlying causes of cognitive dysfunction, we therefore excluded subjects younger than 60 years of age.

We added the following statement to the limitations section on p. XX:

*“[…] due to data availability and increased risk of cognitive deficits being due to neurodegenerative processes, we only included participants over the age of 60, however, accelerated aging starting before this age remained uninvestigated in our study.“*

1. **To disentangle effects of potential atrophy on FDG-PET data and to enable a comparison with the literature (Lee et al., 2022), the PET data should be corrected for partial volume effects (PVE). Of note, the parcellation into 216 regions should results in a number of regions with a very small volume = significant PVE. So, the results both with and without PVE correction should be presented.**

We decided to perform PET analyses without PVE, given that most PET scanners in the clinical setting are PET-CT, rather than PET-MRI scanners, however, an MRI is required to perform PVE. In clinical practice, it is thus unlikely that both, an MRI and FDG-PET scan would be collected for the same individual within a short amount of time. Therefore, to construct clinically relevant brain age estimation frameworks, we treated the two modalities separately. However, to decrease the effect of PVE, we have changed the atlas, used to extract the ROIs from Schaeffer + Tian to AAL1, thereby reducing the number of ROIs from 216 to 116. Brain age accuracy results from the composite (Schaeffer + Tian) atlas have been moved to the supplementary results, but are highly comparable to those obtained from AAL1.

1. **How established and robust is the pipeline that the authors used to calculate BAG? It is user friendly and publicly or commercially available (keeping clinical trials in mind)? How do the results change if e.g., 3- or 6-fold (instead of 5-fold) cross-validation is applied?**

Comparable pipelines as the one used here were also used to estimate brain age from MRI in previous works [5 - 7]. As mentioned in the manuscript, both, support and relevance vector machines have previously been demonstrated to perform brain age estimation well based on small datasets (p. XX, [5]). A five-fold nested cross-validation to perform hyperparameter search and train the models has previously been employed in [6]. Notably, our nested cross-validation procedure proved robustness of results as the range of MAE was considerably small (XX RANGE PET: XX, RANGE MRI: XX). Prior to bias correction, our MAE was XX, which compares well to the MAE ranges of XX – XX reported in [5-7]. The choice of a five-fold cross validation was made, because five-fold is the default setting for conducting (stratified) cross-validation in scikit-learn (see here: <https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.cross_validate.html> and here: <https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedKFold.html>). While we believe the above outlines enough grounds to assume robustness of the model, to convince the reviewer, we added results from 3- and 6-fold cross-validation to the supplementary materials (SM p. XX).

Finally, while we already planned to make our work publicly available on GitHub, we have additionally implemented an easy-to-use pipeline with instructions to allow for BAG assessment of new MRI or FDG-PET scans. All code/pipelines will be made available upon publication, however, a demonstration of the pipeline for external usage has been added to the supplementary materials (SM p. XX). In addition, we added a code availability statement to the manuscript (p. XX):

XX

1. **Why the demographic variables, in particular age, are corrected for by default? BAG is per definition the product of age. Results of the cognitive tests are typically adjusted for the demographic variables (z scores). Superfluous adjustment for variables may lead to spurious associations.**

While composite scores for memory and executive function (ADNI-MEM and ADNI-EF) are indeed z scores, the bias correction procedure may left correlations between BAG and chronological age in some folds of the MCI cohort. These correlations were different, importantly, when BAG was assessed from MRI or FDG-PET. Therefore, we decided to report conservative results by correcting all analyses for age, sex and APOE-e4 carriership to impede the possibility to report false positives. We have added results without correcting for these analyses to the supplementary materials for reference (SM p. XX).

1. **Validation in the OASIS sample is described in Methods, but is absent in Results.**

Validation in the OASIS sample was only mentioned briefly in Table 2 of the results in the previous version of the manuscript. We have now stated these results more explicitly in the flowing text (p. XX):

XX

**References used by reviewers**

- Ebenau JL, Pelkmans W, Verberk IMW, Verfaillie SCJ, van den Bosch KA, van Leeuwenstijn M, Collij LE, Scheltens P, Prins ND, Barkhof F, van Berckel BNM, Teunissen CE, van der Flier WM. Association of CSF, Plasma, and Imaging Markers of Neurodegeneration With Clinical Progression in People With Subjective Cognitive Decline. Neurology. 2022 Mar 29;98(13):e1315-e1326.

- Pike KE, Cavuoto MG, Li L, Wright BJ, Kinsella GJ. Subjective Cognitive Decline: Level of Risk for Future Dementia and Mild Cognitive Impairment, a Meta-Analysis of Longitudinal Studies. Neuropsychol Rev. 2022 Dec;32(4):703-735.

- Scheef L, Spottke A, Daerr M, Joe A, Striepens N, Kölsch H, Popp J, Daamen M, Gorris D, Heneka MT, Boecker H, Biersack HJ, Maier W, Schild HH, Wagner M, Jessen F. Glucose metabolism, gray matter structure, and memory decline in subjective memory impairment. Neurology. 2012 Sep 25;79(13):1332-9.

**References used for rebuttal letter**

[1] Saerens, M., Latinne, P., & Decaestecker, C. (2002). Adjusting the Outputs of a Classifier to New a Priori Probabilities: A Simple Procedure. Neural Computation, 14(1), 21–41. doi:10.1162/089976602753284446

[2] van Harten AC, Wiste HJ, Weigand SD, Mielke MM, Kremers WK, Eichenlaub U, Dyer RB, Algeciras-Schimnich A, Knopman DS, Jack CR Jr, Petersen RC. Detection of Alzheimer's disease amyloid beta 1-42, p-tau, and t-tau assays. Alzheimers Dement. 2022 Apr;18(4):635-644. doi: 10.1002/alz.12406. Epub 2021 Jul 26. PMID: 34310035; PMCID: PMC9249966.

[3] Ranganathan P, Pramesh C, Aggarwal R. Common pitfalls in statistical analysis: Logistic regression. Perspect Clin Res. 2017;8(3):148-151. doi:10.4103/picr.PICR\_87\_17

[4] Jessen F, Amariglio RE, Buckley RF, van der Flier WM, Han Y, Molinuevo JL, Rabin L, Rentz DM, Rodriguez-Gomez O, Saykin AJ, Sikkes SAM, Smart CM, Wolfsgruber S, Wagner M. The characterisation of subjective cognitive decline. Lancet Neurol. 2020 Mar;19(3):271-278. doi: 10.1016/S1474-4422(19)30368-0. Epub 2020 Jan 17. PMID: 31958406; PMCID: PMC7062546.

[5] Beheshti I, Ganaie MA, Paliwal V, Rastogi A, Razzak I, Tanveer M. Predicting Brain Age Using Machine Learning Algorithms: A Comprehensive Evaluation. IEEE J Biomed Heal Informatics. 2022;26(4):1432-1440. doi:10.1109/JBHI.2021.3083187

[6] Baecker, L., Dafflon, J., da Costa, P. F., Garcia-Dias, R., Vieira, S., Scarpazza, C., … Pinaya, W. H. L. (2021). Brain age prediction: A comparison between machine learning models using region- and voxel-based morphometric data. Human Brain Mapping, 42(8), 2332–2346. <https://doi.org/10.1002/hbm.25368>

[7] Beheshti I, Nugent S, Potvin O, Duchesne S. Bias-adjustment in neuroimaging-based brain age frameworks: A robust scheme. NeuroImage Clin. 2019;24:102063. doi:10.1016/j.nicl.2019.102063