We would like to thank the reviewers for their valuable input and overall appreciation of our work. We have revised the manuscript thoroughly. The main changes include: 1) training was now done only on individuals who were cognitively normal, i.e., who had no subjective or clinical indication of cognitive impairment. 2) The focus on our manuscript was shifted more towards describing the cognitive, pathological and prognostic profiles of MRI- or FDG-PET-derived brain age, rather than recommending these brain ages for prognostic purposes, as is reflected in the new title. 3) Accordingly, the logistic regression part of our analyses was extensively edited to put brain age measures into a comparative context with other Alzheimer’s disease biomarkers. 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. In the previous version of the manuscript, we used matching to control for the effects of age and sex. Instead, in the new version of the manuscript, we used the whole cohorts of decliners and stables, but we computed standardized residuals of predictor variables to correct them for not only age and sex, but also years of education and APOE-e4 carriership, as was done in the prior correlation analyses. This is described on p. 7.

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 (pp. 6-7), 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.**

To compare different biomarkers, we now used a priori probabilities of cognitive decline in the training data. This yielded comparable thresholds across all folds (see p. XX).

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 in addition to 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 (pp. 6-7), we now consider the following significance levels: p < .1 = “trend 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. Indeed, research has shown that CSF p-Tau/Aβ1-42 ratio reflects not only tau and amyloid pathology, the two primary hallmarks of AD, but also the risk of impending cognitive decline, well1.

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.**

Rather than providing a classifier for cognitive decline, the purpose of the current study is to explore the two different brain ages yielded from MRI or FDG-PET. Therefore, it is not surprising that not all brain substrates found to be age-associated in our study also hold prognostic potential. However, the re-trained brain age estimation algorithm based on MRI now extensively made use of hippocampal signal, which resulted in strong associations with cognitive performance, AD neuropathology and cognitive outcome. We briefly discuss this on p. 15:

*“Since MRI BAG was mostly estimated from hippocampal volume and predictive of conversion from MCI to dementia, biological processes underlying MRI BAG may be more closely related to AD etiology than those captured by FDG-PET BAG, as it has repeatedly been shown that hippocampal volume plays a pivotal role in memory, and AD-related decline thereof*2*.”*

**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 FDG-PET for the estimation of brain age, as well as to assess the cognitive and neuropathological processes they reflect. It is true, however, that in the previous version of the manuscript, we interpreted our results overly optimistically. In the new version of the manuscript, we refrained from recommending FDG-PET BAG for clinical trials given its inferior performance as a prognostic biomarker. Rather, we suggest that it may, with further validation, be useful as an early diagnostic biomarker of cognitive impairment (p. 14):

*“Our findings suggest that advanced brain age captures brain health, in the form of cognitive and neuropathological variance in the early AD continuum as early as the SCD stage. The observed negative association between FDG-PET BAG and memory performance in the lack of clinically manifest cognitive dysfunction may suggest that FDG-PET BAG has the capacity to detect subtle cognitive decline at its nascent stages. Notably, there was a discernible trend towards higher ME in SCDADNI compared to CNADNI, as well as between declining SCDADNI and stableADNI, although these results did not reach statistical significance. These observations provide preliminary evidence for the utility of FDG-PET BAG as a potential early biomarker for cognitive impairment.”*

Moreover, we recognize that the strong focus on the logistic regression analyses overshadowed the other parts of the manuscript. We re-phrased our aim to “compare the cognitive, pathological and prognostic profiles of brain age estimation from FDG-PET and structural MRI in the early AD continuum” and adapted the title to reflect this alteration.

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 assessed the potential of BAG to predict cognitive outcome. Following the author’s suggestion, we have investigated the prognostic value hippocampal volume, SUVR in precuneus, chronological age for reference, as well as other established biomarkers of AD using the same pipeline we used to assessed how well BAG predicts cognitive outcome. In the revised version of the manuscript, we were able to show that

1. MRI brain age, when trained solely on CN, is very, but not uniquely, dependent on hippocampal signal. This intriguingly suggests that AD-related neurodegeneration can be considered a form of accelerated brain aging on MRI (p. 15).
2. MRI BAG (AUC = .73), which was derived from multi-dimensional data predicted cognitive outcome comparably well as established AD biomarkers (AUC = .70 - .78) after correcting for age, sex, years of education and APOE-e4 carriership. Based on these results, we recommend further testing of MRI BAG as a complement measure for the development of reliable prognostic tools for MCI-to-AD conversion (p. 14).
3. **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 limitations 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 = 185) 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 were therefore discarded from our sample. Amyloid and APOE-e4 status are presented in Table 1 (p. 21).

Unfortunately, due to the small sample size of decliners in the CN (n=16) and SCD cohorts (n=10), we could not reliably separate these groups for the prediction of cognitive outcome analyses. Hence, while we isolated CN and SCD for the training of our models, as well as to compute correlations of BAG with cognitive performance and AD neuropathology, CN and SCD were grouped to a “cognitively unimpaired” (CU) cohort for the last part of our investigations. The proportion of decliners was comparable in the CN (11%) and SCD group (12%). This justification for grouping together CN and SCD is described on p. 7:

*“Due to the small number of decliners in the CNADNI (n=16≙10%) and SCDADNI samples (n=10≙12%), we combined the two groups to a cognitively unimpaired (CUADNI) cohort.”*

**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 hope that with the new focus of our paper, this has become more apparent.

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. 15:

*“[…] 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 investigate the profiles of clinically realistic 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 and serve to prove robustness of our method.

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 works3–5. As mentioned in the manuscript, both, support and relevance vector machines have previously been demonstrated3 to perform brain age estimation well based on small datasets (p. 4). A five-fold nested cross-validation to perform hyperparameter search and train the models has previously been employed in Baecker et al.5. Notably, our nested cross-validation procedure proved robustness of results as the range of MAE in the test set was considerably small (range MRI: [2.00, 2.73], range PET: [2.20, 2.98]) and consistent with results obtained in the paper we used for reference for the bias adjustment procedure (MAE = 2.36 years)4. Prior to bias correction, our MAE was in the range of [4.12, 4.52] for MRI models and [3.95, 4.38] for FDG-PET-derived models, which compares well to the average MAE of 4.21 and 3.43 reported for MRI- and FDG-PET-based models, respectively, prior to bias adjustment in Lee et al.6. 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>).

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 in order to enable researchers to quickly access the tools we developed. 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. 17):

*“The code used for this project will be made publicly available on the GitHub page of the first author upon publication.”*

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).

Added explanation: Correction for age was added despite successful bias correction given the age-associativity of the dependent variables.

Variables we corrected for were also used in ADNI-MEM paper.

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