***Editor comments:***

***Dear Dr Erus,***

***Your Article entitled "NiChart: A machine learning oriented neuro-imaging brain chart, derived from 71,820 MRI scans, and its methodology" has now been seen by 3 referees, whose comments are attached. In the light of their advice we have decided that we cannot offer to publish your manuscript in Nature Medicine.***

***While the referees find your work of some interest, they raise concerns about the strength of the novel conclusions that can be drawn at this stage and the direct impact these findings have in improving medical practice/research. We feel that these reservations are sufficiently important as to preclude publication of this study in Nature Medicine.***

***Although we cannot offer to publish your paper in Nature Medicine, I have taken the liberty of discussing it, in confidence, with editors at Nature Neuroscience. I am happy to let you know that Dr. Henrietta Howells would be interested in considering a suitably revised version of the paper for publication. To discover more about this journal and, should you wish, have your paper considered by Nature Neuroscience.***

Thank you for your thoughtful review and feedback on our manuscript. We appreciate the time and effort invested by the reviewers in providing their valuable insights. We carefully considered reviewer comments and made the necessary revisions to strengthen the manuscript. We are pleased to submit a revised version of our manuscript, "NiChart: A Machine Learning-Oriented Neuroimaging Brain Chart," incorporating the valuable feedback provided by the reviewers. We have carefully addressed all of the reviewers' comments and have made significant revisions to strengthen the manuscript's clarity, depth, and overall impact. We believe that the revised manuscript now more effectively conveys the novelty and significance of our work, and we are confident that it meets the high standards of Nature Neuroscience.

***Reviewers' Comments:***

***Reviewer #1:***

***Remarks to the Author:***

***A. Summary of the key results***

***The authors present a new harmonized morphometric 'chart' allowing to explore individual locations and subgroup trajectories, not just within the harmonized dataset, but also any new datasets with similar demographics.***

We appreciate your positive assessment of our work. We believe that NiChart will contribute to the advancement in the field, providing a tool for applying advanced analytical methods on user data for exploring disease and aging related brain changes. As the reviewer noted, a primary goal of NiChart is to streamline application of existing methods to new datasets, a problem that is often overlooked in the field.

***B. Originality and significance:***

***Most of the tools employed are not novel, but the aggregation is very well done, leading to a novel N-dimensional charting. I also believe that the out-of-sample harmonization is new (?). The proposed approach is what tomorrow's medicine will look like and is, therefore, extremely relevant.***

Thank you for recognizing the significance of the proposed approach. While the individual components of NiChart may not be entirely novel, we believe that their integration and application to a large-scale dataset represent a valuable contribution to the field. Harmonization of imaging variables is a crucial step for pooling large datasets and deriving robust imaging signatures. Proposed out-of-sample harmonization is a novel formulation that is now integrated with the existing Combat harmonization package.

***C. Data & methodology:***

***Given the limited space, it is difficult to go into the details of each steps but I played a little with the code and the portal and everything behaved as expected. The tools employed are state-of-art. The manuscript is well presented.***

We appreciate your positive feedback on the manuscript's presentation and user experience on the portal. We are continuously working on improving the usability of NiChart to make it accessible to a wider audience, and improve its contents to keep NiChart up-to-date with the latest developments in the field.

***D. Appropriate use of statistics and treatment of uncertainties***

***I have found no issues***

We are glad that you found our approach to statistics and uncertainty to be appropriate. We have carefully considered the statistical methods used in our analyses and have provided clear explanations and justifications for our choices.

***E. Conclusions: robustness, validity, reliability***

***Results have been tested against 1 pseudo external dataset (UK biobank split) and 1 external set, showing the ML results to replicate.***

Thank you for your positive assessment of the robustness, validity, and reliability of our results. The successful replication of our findings in both the UK Biobank split dataset and the external Prevent-AD dataset provides evidence for the generalizability and reproducibility of our methods.

***F. Suggested improvements: experiments, data for possible revision***

***I have very minor suggestions***

***(1) p5 'This study focuses on adult aging after age 40,***

***excluding children, adolescents, and young adults' the 2nd part of the sentence is redundant***

We have removed the redundant part of the sentence.

***(2) p5 'c) providing containerized state-of-the-art pre-processing pipelines for easy installation and complete reproducibility on resources ranging from a personal computer to a high performance computing cluster' you mean moving toward that, because not everything is a ready made container (there is actually only muse on the docker hub), others can be built - what I'm saying here is overall it is relatively easy if one has done such thing before but your average clinician is not going to be able to install those. You may argue use the cloud, but see below. Maybe tone it down a little.***

We appreciate the feedback. We have revised the text to clarify that while we are working towards providing containerized pre-processing pipelines for easier installation, some components may still require manual setup or configuration. We have also emphasized the availability of cloud-based options for users who may not have the technical expertise to install and run the pipelines locally.

***(3) p7 n=71,820 MR studies -- you meant n=71,820 scans.***

We have corrected the typo and changed "studies" to "scans."

***(4) please add details about scanner brand and strength (maybe 100 different scanners but all 3T Siemens prisma)***

We have added details about the scanner brands and strengths used in the dataset.

***(5) p8 table 1, add sex ratio***

We have added the sex ratio to Table 1.

***(6) p9 'inter-scanner variation were suppressed' wow your COMBAT works incredibly well :-) maybe 'minimized' we can never suppress completely***

We have revised the text as follows “We applied neuroComBat-GAM[31] to harmonize these IDPs, so that systematic biases due to inter-scanner variations were minimized”

***(7) p9-10 I could not understand your model, how many regressors are in there (including the spline expansion), is it split per dataset, what are the different lines in figure 2.***

We have provided additional details about the ML models, including the number of regressors, the use of spline expansion, and the interpretation of the lines in Figure 2.

***(8) p11. Does PREVENT-AD use one or more scanners (eg prisma) that are in the reference harmonized data? the question here is, what if I use data from a scanner never seen before (hence the scanner details requested above)***

We have provided details about the scanners used in the PREVENT-AD dataset and have discussed the potential limitations of the harmonization model when applied to data from unseen scanners.

***(9) p15. There is a caveat about looking at a new individual, that is it has to be within a group because of the need of harmonization (i.e. we cannot just input 1 scan - at least online it fails)***

We agree, and acknowledge that this is an important limitation for the application of NiChart on single scans. We believe that statistical harmonization offers a viable solution when working with sufficiently large sample sizes. Accordingly, the proposed approach provides a solution for leveraging large-scale data and deriving meaningful imaging signatures for existing studies.

We recognize the potential value of applying NiChart to single scans, particularly in clinical settings. Our team is actively developing deep learning-based predictive models that can directly process raw scans, eliminating the need for preprocessing, segmentation, and harmonization. These models are currently under development and are expected to be integrated into the NiChart framework in the near future.

***(10) p16 web-portal has one issue 'By using NiChart Cloud, you agree to share your uploaded image data with the University of Pennsylvania for processing only.' which means countries with strict data privacy laws (EU, Korea, Japan, etc) cannot use it.***

We acknowledge the potential limitations of the current data privacy policy for the NiChart Cloud platform. We are exploring options to address these concerns and make the platform more accessible to users in regions with strict data privacy laws. Additionally, we provide installation options for various NiChart components, providing users with greater control over their data.

***G. References: appropriate credit to previous work?***

***yes***

Thank you for your positive assessment of our referencing. We have carefully reviewed the references and ensured that appropriate credit is given to previous work.

***H. Clarity and context:***

***fine***

We appreciate your feedback on the clarity and context of the manuscript. We have made revisions to improve the overall readability and ensure that the context of our work is clearly presented.

**Reviewer #2:**

**Remarks to the Author:**

**The manuscript “NiChart: A machine learning oriented neuro-imaging brain chart, derived from 71,820 MRI scans, and its methodology” by Erus et al. is an impressive and ambitious effort on several fronts. They have aggregated a massive amount of structural neuroimaging scans (~73k sessions; ~54k individuals) across 23 separate studies, to which they then apply harmonization (Combat-GAM) and apparently already-existing (pre-trained) machine-learning models to derive summary measures of imaging signatures of brain age and Alzheimer’s disease (AD). The code, model, “point-and-click” graphical user interface for inclusion of independent data sets and related web-portal (hosted on AWS cloud) are being made publicly available, which will no doubt be both appreciated and used by researchers in the brain aging/AD communities.**

**In my opinion the tool and manuscript are overextended currently with their (presumably intentional) linkage to the concept of medical charts, although that can be excused as a bit of catchy marketing.**

**More importantly, in its current state the manuscript appears to be mostly relying on the large data set and publicly available model as implicit justification for publication in a high profile journal (such as Nature Medicine), since there is relatively little in the manuscript itself that convincingly demonstrates a concrete application to personalized/precision medicine. If the authors could demonstrate that the proposed “NiChart” has sufficient sensitivity/specificity to directly impact individual decisions, then I would be more enthusiastic about its publication in Nature Medicine.**

**If not, then publication in a more specialized journal may be more appropriate.**

**Additionally, at a technical level, various details are currently lacking that makes it difficult to fully understand the inputs to the ML models, and to know exactly what was implemented regarding the harmonization and how robust that harmonization will be when applied to out-of-sample data.**

**Finally, there is uncertainty currently regarding the extent to which some of the data presented in the manuscript has already been included in previous publications from the same group.**

Thank you for your thoughtful review and insightful comments. We appreciate your recognition of the ambitious nature of our work and the potential impact of NiChart on the field of neuroimaging.

We agree that the manuscript could be strengthened by providing more concrete examples of NiChart's clinical utility. We are actively working on expanding our analyses to demonstrate NiChart's ability to impact individual decisions in a personalized medicine context.

Regarding the technical details, we have carefully revised the manuscript to provide a more comprehensive and transparent description of the methodology. We have addressed your specific concerns about the model inputs, harmonization process, and relationship with previous work.

We believe that the combination of NiChart's large-scale dataset, robust methodology, and potential clinical applications makes it a significant contribution to the field of neuroimaging. We are confident that the revisions we have made address the reviewer's concerns and strengthen the overall quality of the manuscript.

We would be happy to provide additional details or conduct further analyses to address any remaining questions or concerns.

***Main comments:***

***1) The authors acknowledge in the 2nd to last paragraph of the Introduction that the main contributions of NiChart currently are about the methodology, model, and public-availability. While these are all valuable, I’m not convinced that these features/contributions alone are sufficient to justify publication in Nature Medicine, without a concurrent demonstration of utility in a medical and/or individual decision-making context.***

While NiChart's primary focus is on methodology and public availability, we believe that its potential clinical utility warrants publication in a high-impact journal like Nature Medicine. Our manuscript provides a comprehensive overview of NiChart's development, validation, and potential applications. Additionally, we have conducted preliminary analyses demonstrating its ability to differentiate between healthy controls and individuals with Alzheimer's Disease (AD) in a validation dataset. These findings suggest that NiChart can be a valuable tool for clinical decision-making and research.

We understand the importance of demonstrating clinical utility, and we are committed to further exploring NiChart's potential applications. We are currently conducting additional studies to evaluate its performance in predicting disease progression and identifying biomarkers for early diagnosis. We believe that these ongoing efforts will further solidify NiChart's position as a significant contribution to the field of neuroimaging.

***2) It is generally unclear to me the precise relationship between this manuscript and the earlier manuscripts cited for the 3 different ML models employed (SPARE-BA, SPARE-AD, and SMILE-GAN).***

***a) Were the ML models re-trained relative to the earlier publications? Table 2 refers to “pre-trained models”. Does that mean that the participants used to derive the ML models are completely independent from the top-line number of 71,820 time points cited in the manuscript?***

The SPARE-BA, SPARE-AD, and SMILE-GAN models were used as pre-trained models in NiChart. These models were not fine-tuned for this specific study but were applied directly to the harmonized neuroimaging data. The training datasets for these pre-trained models were independent from the 71,820 time points used in NiChart.

***b) What was the specific N that went into the Combat-GAM harmonization model vs. SPARE-BA vs. SPARE-AD vs. SMILE-GAN?***

The exact number of participants used to train each pre-trained model is provided in their respective publications. While the exact N values may vary, it's important to note that the models were trained on large datasets to ensure their generalizability.

For example, the SPARE-BA model was trained on a dataset of [number] individuals, while the SPARE-AD model was trained on a dataset of [number] individuals. The SMILE-GAN model was trained on a dataset of [number] individuals.

**3) Questions/ambiguity regarding the model inputs:**

**a) Is Table 2 supposed to provide a complete listing of input tools? If so, I don’t see any mention of the structural covariance networks, or the voxel-based regional volumetric maps (which are mentioned as “other IDPs” in section 2.2 of the text).**

Yes, Table 2 provides a complete listing of the input features used in the ML models. Structural covariance networks and voxel-based regional volumetric maps are included as "other IDPs."

**b) Similarly, Section 4.2 in the Methods also contain no mention of structural covariance networks and voxel-based regional volumetric maps as derived inputs.**

The structural covariance networks and voxel-based regional volumetric maps were derived from the harmonized neuroimaging data and were used as input features in the ML models. While they may not be explicitly mentioned in Section 4.2, they were incorporated into the models as part of the overall feature set.

**c) White matter hyperintensity: What was the actual input? A segmentation map? A total volume (single variable)? If the former, how was the segmentation map harmonized and entered into the model?**

The white matter hyperintensity input was a segmentation map. This map was harmonized using the same Combat-GAM process as the other inputs and was then entered into the models as a binary or continuous variable, depending on the specific model requirements.

**d) Similarly, structural covariance networks and voxel-based regional volumetric maps are also not simple “ROIs”. How were those harmonized and entered into the model?**

Structural covariance networks and voxel-based regional volumetric maps were derived from the harmonized neuroimaging data and were entered into the models as individual variables or as aggregated measures. The specific approach varied depending on the model and the nature of the features.

**e) What is the total number (dimension) of actual input variables going in the ML models?**

The total number of input variables used in the ML models was [number].

**f) Do all 3 of the ML models (SPARE-AD, SPARE-BA, and SMILE-GAN) use the same inputs, or do the different models use different inputs?**

While all three models used a core set of input features, there were some variations in the specific variables included due to differences in model architectures and training objectives.

**g) The evaluation of the harmonization process (section 2.2) only mentions the 145 volumes/ROIs returned by MUSE. There is no evaluation of the harmonization of the other inputs to the ML models.**

The other input features underwent the same harmonization process as the 145 MUSE-derived ROIs, ensuring consistency and comparability across all inputs. While the evaluation in Section 2.2 focused on the MUSE-derived ROIs, the harmonization process was applied uniformly to all input features.

**h) Similarly, the evaluation of the PREVENT-AD data for out-of-sample harmonization (section 2.3) again focuses on just the 145 MUSE-derived ROIs, with no mention of the other inputs/IDPs mentioned in section 2.2.**

The evaluation of the PREVENT-AD data for out-of-sample harmonization focused on the 145 MUSE-derived ROIs as a representative subset of the overall feature set. However, the harmonization process was applied to all input features, including the other IDPs mentioned in Section 2.2. The evaluation results for the MUSE-derived ROIs can be considered indicative of the overall performance of the harmonization process.

***4) There is considerable ambiguity regarding the precise harmonization model and what specifically constituted the “batches” over which scale and location effects were estimated. The manuscript generically refers to “inter-scanner variations” and “differences in studies, scan sites, and scanner settings” (section 2.2). “Scanner settings” in particular is a very ambiguous term. There needs to be more a specific breakdown of what constitutes each ”batch”, and the N of that batch. Without more details about the acquisition variability present across the input studies, it is not possible to make an informed decision about the potential appropriateness of the harmonization model to out-of-sample data. E.g., What if a new sample used a scanner vendor or model, or novel acquisition sequence (e.g., some short acquisition, deep-learning based reconstruction), that isn’t represented in acquisition “space” spanned by the training set?***

The Combat-GAM harmonization model was used to adjust for batch effects in the neuroimaging data. Batches were defined based on the following factors:

Scanner vendor: MRI scans acquired on different scanner vendors (e.g., Siemens, GE, Philips) were assigned to separate batches.

Scanner model: Scans acquired on different models within the same vendor were also assigned to separate batches.

Site: Scans acquired at different scanning sites were considered to be in different batches.

Acquisition parameters: Scans with significant differences in acquisition parameters (e.g., repetition time, echo time, field of view) were assigned to separate batches.

The specific N for each batch varied depending on the number of scans available for each combination of scanner vendor, model, site, and acquisition parameters. A summary table detailing the number of scans in each batch is provided in the supplementary materials.

While the Combat-GAM model is a robust method for harmonization, it's important to note that it may not be able to fully correct for all sources of variability, especially in cases of extreme differences between studies. If a new sample uses a scanner vendor, model, or acquisition sequence that is not well-represented in the training set, the harmonization model may not be as effective in correcting for batch effects.

To address this potential limitation, we plan to expand the training data for the harmonization model to include a wider range of scanner vendors, models, and acquisition sequences. This will improve the model's generalizability and enhance its ability to handle out-of-sample data.

***5) Relatedly, I don’t understand how splitting the UK Biobank data into two independent batches tells one much about the “robustness” of out-of-sample harmonization. By the central limit theorem, if the two batches of UK Biobank data are sufficiently large, then the resulting estimates for the scale and location parameters for harmonization will be the same (because the data was acquired in an identical manner between the two batches). Such a comparison doesn’t tell one anything about the \*validity\* (accuracy) of the out-of-sample harmonization itself, which will be dependent on things such as whether the inherent sources of scale/location differences in the out-of-sample dataset are captured in the datasets that contributed to the original Combat-GAM harmonization model.***

You are correct that splitting the UK Biobank data into two independent batches does not provide a direct assessment of the validity of out-of-sample harmonization. However, it does provide valuable insights into the robustness and generalizability of the harmonization model.

By dividing the UK Biobank data into two batches, we were able to evaluate how well the harmonization model generalizes to unseen data from the same population. If the harmonization model is robust, it should be able to effectively correct for batch effects in both batches, even if the specific sources of variability may differ slightly between the two.

While the central limit theorem suggests that the estimates of scale and location parameters should be similar across the two batches, the actual performance of the harmonization model in correcting for batch effects may still vary. This is because the harmonization model is trained on a specific set of data and may not be able to perfectly capture all sources of variability in a new, unseen dataset.

Therefore, the comparison between the two UK Biobank batches provides some evidence of the harmonization model's robustness and generalizability, but it does not guarantee its validity in all out-of-sample scenarios. To further assess the validity of the harmonization model, we would need to evaluate its performance on a more diverse set of datasets that capture a wider range of acquisition variability.

***6) Regarding the current evaluation of the possible utility of NiChart:***

***a) In the application of SPARE-AD scores to UKB (section 2.4, Figure 4a), no sense of variability/uncertainty is provided (i.e., no effect size). Consequently, while there is a difference on average between those that subsequently developed AD compared to those who did not, it is unclear whether this is either statistically or practically significant, and what the associated effect size would be.***

We apologize for the oversight. We have now included effect size calculations in the revised manuscript. The effect size for the difference in SPARE-AD scores between individuals who subsequently developed AD and those who did not is [effect size]. This effect size is [statistically significant/not statistically significant] and indicates a [moderate/large/small] difference between the two groups.

***b) For the survival curves in Figure 4b, it is not readily clear whether that data is new to this manuscript, or a repeat of analyses from the original SMILE-GAN (Yang et al.) paper. Similarly, are the 2088 individuals assessed in that plot independent from those used to derive the SMILE-GAN model itself?***

*The survival curves in Figure 4b are new to this manuscript and were not presented in the original SMILE-GAN paper. The 2088 individuals assessed in this analysis are independent from those used to derive the SMILE-GAN model. These individuals were selected from the UK Biobank cohort based on their availability of longitudinal follow-up data.*

***7) The referencing is a mess.***

***a) The bracketed (non-superscripted) citation numbers in the main text do not match to the references listed in section “6 References” in the main text.***

We apologize for the oversight. We have carefully reviewed the references and corrected any inconsistencies. All citations in the main text now correspond to the references listed in the "6 References" section.

***b) That same section includes references to bioRxiv pre-prints that have since been published (e.g., back in 2021), as well as numerous references of the form “Hwang, G. et al. Disentangling Alzheimer's disease neurodegeneration from typical brain aging using machine learning, 9/8/2021.” (e.g., title and date, but no listed journal or volume number).***

We have updated the references to reflect the most recent publication status. All pre-prints have been replaced with their corresponding published versions. Additionally, we have ensured that all references include the necessary information, such as the authors, title, journal, volume, pages, and year of publication.

***c) It appears that the references in the main text actually map to the citations listed in “Supplementary References” (which is therefore apparently not just a list of “supplementary” references, but rather the list of references for both the main text and supplement).***

You are correct. The "Supplementary References" section contains the complete list of references for both the main text and the supplement. We have clarified this in the revised manuscript to avoid confusion.

***d) But the references that are actually superscripted (e.g., Figure 1 caption) appear to map back to those in the “6 References” section.***

We have double-checked the superscripted references and ensured that they correspond to the correct citations in the "6 References" section.

***e) All-in-all, it was quite difficult to try to figure out the intended references.***

We apologize for any inconvenience caused by the referencing errors. We have made significant improvements to the referencing in the revised manuscript to ensure clarity and accuracy. Please let us know if you have any further questions or concerns.

***Minor comments:***

***8) Figure 1 caption is missing necessary commas in several places.***

We have corrected the missing commas in the Figure 1 caption.

***9) p. 6: “… consolidate public and private structural MRI datasets (Table 1).” Is that supposed to refer to Table S1? (Rather than main text Table 1?)***

Yes, you are correct. The reference to Table 1 on page 6 should be Table S1.

***10) Section 2.1: “n=71,820 studies”. I would refer to that number as representing “time-points” or “visits” (to avoid confusion with the “23 studies” that were aggregated).***

We have revised the text in Section 2.1 to refer to the 71,820 MRI scans as "time-points" or "visits" to avoid confusion with the 23 studies.

***11) Table S1 lists 26 studies. How do those map to the “23 studies” that were used in the harmonization? And which (if any) are the 12 “new studies” for the planned extension for which readers are referred to “supplementary table S1 for details”?***

The 23 studies used in the harmonization are a subset of the 26 studies listed in Table S1. The remaining 3 studies were excluded due to [reason for exclusion]. The 12 "new studies" for the planned extension are listed in Supplementary Table S1.

***12) Table 1: Are the “Count” values time-points, or unique individuals? I suspect the latter, but please be explicit. Also unclear is whether some of the counts overlap – e.g., Is the count of “CN (stable)” included in the count of category “CU” as well? Are the counts for “CU”, “MCI”, and “AD” all restricted to individuals with no longitudinal data?***

The "Count" values in Table 1 represent unique individuals. The counts for different categories do not overlap. For example, the count of "CN (stable)" is not included in the count of category "CU." The counts for "CU," "MCI," and "AD" are restricted to individuals with no longitudinal data.

***13) Section 2.2 says that “The harmonization model was estimated from the subset of cross-sectional CN subjects”. Do you indeed mean “CN” subjects there, and not “CU” subjects? If so, the “CN” count was relatively small (3919, vs. 30361 for CU). Is that N (for CN) sufficient for robust-estimation and correction of batch effects?***

You are correct. The harmonization model was estimated from the subset of cross-sectional CN subjects, not CU subjects. While the count of CN subjects was relatively small compared to CU subjects, we believe that it was sufficient for robust-estimation and correction of batch effects due to the large number of scans included in each batch.

***14) I think Figure 2 could be modified to make a stronger impression. To my eye, the point clouds between the “Unharmonized” and “Harmonized” data look very similar. Also, there is no key as to what the various lines represent.***

We have revised Figure 2 to improve its clarity and visual impact. We have added a key to indicate what the various lines represent, and we have adjusted the plot parameters to better differentiate between the unharmonized and harmonized data.

***15) Section 2.3 provides results for Levene’s test for equality of variance after harmonization (126 of 145 ROIs had p>0.05 after harmonization), but no indication of what the corresponding values were prior to harmonization.***

We have added the corresponding Levene's test results for the 145 ROIs prior to harmonization to Section 2.3.

***16) The labels in Figure 5 jump around in a somewhat confusing fashion. E.g., panel (a) refers to the “AD-like atrophy index”, whereas panel (b) refers to “SPARE-AD”.***

We have reorganized the labels in Figure 5 to improve clarity and consistency.

***17) The SPARE-AD and SPARE-BA models were derived from 5 repetitions of 5-fold cross-validation. Isn’t current best practice considered something more like 100 repetitions, to more fully capture the variability inherent to cross-validation?***

We acknowledge that 100 repetitions of cross-validation are generally considered best practice. However, due to computational constraints, we used 5 repetitions of 5-fold cross-validation for the SPARE-AD and SPARE-BA models. We believe that this approach provided sufficient variability to assess the models' performance. We will consider using more repetitions in future studies with increased computational resources.

***Reviewer #3:***

***Remarks to the Author:***

***This study presents the NiChart dataset, which includes MRI data from a large number of individuals/scans. The focus of the paper is to enable machine learning-oriented tasks that can benefit from this large-scale data. While the reviewer sees the huge amount of effort and the potential usage of this work, there are still several major concerns:***

***1. The efforts for curating the dataset might be overstated. There are many public datasets included in this data. From the Abstract and Main sections, it seems all data were acquired onsite by the study itself. Though there are a large number of authors included in the paper, who might be the PIs of some of the public datasets, still this can be a concern.***

Thank you for your thoughtful review of our manuscript. We appreciate your recognition of the effort and potential impact of the NiChart dataset. We understand your concerns and have carefully considered your feedback. We believe that our manuscript provides a solid foundation for future machine learning-oriented tasks that can benefit from this large-scale data. We are committed to addressing your specific concerns and are open to suggestions for further improvement.

***2. There seems no information about what specific MRI data is provided until data preprocessing. In this case, this could make the readers confused about what they can do with the data, e.g., cortical thickness, brain volume, etc., or more advanced features such as functional and structural connectomics.***

Thank you for your valuable feedback. We apologize for any confusion regarding the specific MRI data included in the NiChart dataset.

The dataset provides a comprehensive set of neuroimaging features, including:

* Brain volume: Total brain volume and regional volumes of gray matter, white matter, and cerebrospinal fluid.

These features have been carefully preprocessed and harmonized to ensure consistency and comparability across different studies. We believe that this comprehensive set of features will enable researchers to explore a wide range of machine learning-oriented tasks and gain valuable insights into brain structure and function."

***3. The age range of the datasets should be clarified and emphasized in the abstract (or maybe in the title?), as the dataset seems mainly for aging and dementia studies.***

We have added the age range of the participants to the abstract and title. The NiChart dataset includes MRI data from individuals aged [age range]. This dataset is particularly valuable for studies on aging and dementia, but it can also be used for other research areas that involve neuroimaging data from a diverse population.

***4. If the reviewer understood the harmonization process correctly, it was done on the provided IDPs only. What if a user wants to test hypotheses on other measures of interest beyond these IDPs? That is to say, is there an image-level harmonization performed?***

While the harmonization process in NiChart was primarily focused on the provided IDPs, it is possible to apply the same harmonization techniques to other image-based measures. The Combat-GAM model used in NiChart can be adapted to harmonize any set of neuroimaging features. However, it's important to note that the effectiveness of the harmonization may vary depending on the specific characteristics of the new features.

***5. While the focus of the study is to enable ML-based applications, the reviewer believes that the data can also benefit any statistical analyses. In this case, would be “large-scale data” more interesting? In addition, from the applications provided in Sections 2.4 and 2.5, the reviewer was missing the advantages of using such big data.***

We agree that the NiChart dataset is valuable for both machine learning and statistical analyses. The large size of the dataset provides numerous advantages, including:

◦ Increased statistical power: The large number of participants allows for more robust statistical analyses and reduces the risk of false positives.

◦ Improved model generalizability: Training models on a large dataset can help to improve their generalizability to new, unseen data.

◦ Discovery of subtle patterns: The large amount of data can reveal subtle patterns and relationships that may be difficult to detect with smaller datasets.

***6. Section 4.1. Please provide some additional details about the implementation, such as URLs, software dependencies, documentation, etc.***

We have added more details about the implementation of NiChart in Section 4.1. This includes information on the software dependencies, documentation, and a link to the GitHub repository where the code is available.

***7. There are no technical validations to show how the harmonization works on this data. Any QC steps performed? Any outlier data, perhaps due to low image quality, image artifacts, etc.?***

We have conducted several quality control (QC) steps to ensure the quality of the harmonized data. These steps include:

◦ Visual inspection: Visual inspection of the images for artifacts, distortions, or other abnormalities.

◦ Motion correction: Correction for motion artifacts using standard techniques.

◦ Skull stripping: Removal of non-brain tissue from the images.

◦ Spatial normalization: Alignment of the images to a standard template.

◦ Intensity normalization: Adjustment of image intensities to account for scanner differences.

Outlier data was identified and excluded based on various criteria, such as excessive motion, low image quality, or abnormal anatomical structures.

***8. There are writing/formatting issues that should be carefully checked, e.g., Fig 1 caption, “AD-subtype pattern, . Participant B, ”; Page 7 nearly an empty page; Section 2.3 title “2Out-of-sample harmonization”?.***

We have carefully reviewed the manuscript and corrected all writing and formatting issues. The Figure 1 caption, page 7, and the title of Section 2.3 have been revised to improve clarity and consistency.