PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	The Arts for Ageing Well: A Propensity Score Matching Analysis of the Effects of Arts Engagements on Holistic Wellbeing among Older Asian Adults above 50 years of age
AUTHORS	Ho, Andy H Y; Ma, Stephanie Hilary Xinyi; Ho, Moon-Ho Ringo; Pang, Joyce Shu Min; Ortega, Emily; Bajpai, Ram

VERSION 1 – REVIEW

REVIEWER	Emma Wood School of Events, Tourism and Hospitality Management Leeds Beckett University
REVIEW RETURNED	13-Feb-2019

GENERAL COMMENTS	A well executed study that adds to the growing evidence base in this area Very well written in an accessible style but includes a few typos. A further proofreading needed Consistency needed in the abstract and earlier sections in how the sampled population are referred to. ie not 'elders' or 'elderly' or 'aged' or 'older' - better to use 'people over 50 years of age' as other terms subjective and relative. Demographic data might usefully have included household composition as living alone likely to be significant factor in subjective wellbeing. If this is identified from the interpersonal support scale it would be good to have this explained The discussion if findings and conclusions needs to include a commentary on the direction of the causal relationship for several of the significant results. For example, a sense of belonging may be a precursor to art activity rather than a result of it. The implications for this would be that an enabler of arts participation is to first facilitate belonging. The same holds true for several of the other wellbeing, QoS measures.

REVIEWER	Helen Chatterjee University College London UK
REVIEW RETURNED	26-Feb-2019

GENERAL COMMENTS	This as an interesting population sampling approach using observational and PSM analysis to address the value of creative ageing in Singapore. There are many positive aspects to the study - not least the fact that this is the first comprehensive study based

	on an Asian population sample, but the paper would benefit from drawing on a wider range of literature from the extensive evidence base. Overall the manuscript is well written, clear and informative; the study adds value to the existing evidence base by using self-report measures to correlate with past arts engagement across a small, but significant, population sample. See specific comments below: Some of the language is overly subjective, for example in paragraph 1 of the Introduction older adults are described as being 'agonized by depression and dementia' and in places overstates the findings. In addition to References 6 and 7 in paragraph 2 of the Introduction I would recommend citing the UK's Creative Health report which brings together over 1000 peer-reviewed and grey literatures on arts and health; this is the most extensive review on the benefits of arts to health and the authors would no doubt benefit from integrating its findings in the present study. (Creative Health: The Arts for Health and Wellbeing. Inquiry Report. 2017. https://www.artshealthandwellbeing.org.uk/appg-inquiry/). The introduction (para 3) suggests that a literature search has been conducted on creative ageing and health, but this overview fails to mention numerous arts and older adult studies - see below to name but a few - the Creative Health report highlights many others: Coulton, S., Clift, S., Skingley, A. and Rodriguez, J. (2015) Community singing and health in the older population: A randomised controlled trial, British Journal of Psychiatry, 211, 6, 1–6. Thomson, L. J., Lockyer, B., Camic, P. M. & Chatterjee, H. J. (2017). Effects of a museum-based social prescription intervention on quantitative measures of psychological wellbeing in older adults. Perspectives in Public Health. DOI: https://doi.org/10.1177/1757913917737563 Kattenstroth, J-C., Kalisch, T., Holt, S., Tegenthoff, M. & Dinse, H. R. (2013). Six Months of Dance Intervention Enhances Postural Sensorimotor and Cognitive Performance in Elderly without Affecting Cardio-respiratory Functions. Frontiers in Aging Neuroscience, 5 (5). In Methods the authors should define more clearly what they mean by ‘active’ and ‘passive’ engagement for each of the 8 art forms. The Results compared findings with a control group but this is not mentioned in the Study Design; the Methods should include an explanation of the control group and how these data were sourced. Further information is needed about how participant health status was assessed. The Discussion presents a useful synthesis of some of the mechanisms underpinning the value of arts engagement to health which is very useful, but it would be valuable to highlight the range of variation in the various art forms involved and contrast this with previous studies. There should also be recognition of the limitations of the study, including the use of incentives to encourage participation and the

	accuracy of memory recall for activities undertaken at least 3-6 months prior to survey data collection.

REVIEWER	Bruno Arpino University of Florence, Italy
REVIEW RETURNED	12-May-2019

GENERAL COMMENTS	This is an interesting study that about the effect of active and passive engagement on several measures of wellbeing of older adults in Singapore. Abstract All reported associations between art engagement and wellbeing outcome measures are positive but t statistics are negative. Please, double check. Methods Please clarify whether or not all individuals age 50+ were interviewed in each randomly selected household. One important confounding factor is participation in activities different than art. The authors collected data on physical activity but it is not clear whether authors collected data on social activities, such as volunteering and participating in a (non art) club or association. If information on other activities was not collected, the authors should justify this choice and acknowledge in the paper limitation that such an important confounding factor may bias the estimates. For example, individuals involved in art activities may also be involved in volunteering and the effect on wellbeing may simply be the effect of engagement in general and not of art engagement per se. Related to the previous comment, it is not clear actually whether involvement in other activities is part of some outcome measures. On page 6 the authors mention “daily activities” and “social functioning” as components of the first two outcome measures, respectively. Thus, it seems that a potential confounding factors is in fact considered as part of the outcome. The authors should clarify the role of these other activities and social functioning (should they be considered as confounders or as outcomes?) and also how the different components of the outcome measures are defined and measured. The authors should provide much more detail on the implementation of the PSM. What is the estimand of interest (ATE or ATT)? Did the author use matching with or without replacement? What balance diagnostic was considered? I guess it was the standardized mean difference between treatment and control group but the authors should be more explicit on this point. They should also check table 3: before matching it seems that they show the unstandardized mean difference. Moreover, the reported values of standardized difference do not seem to match with those reported in the text (e.g., median biases of 1.4% and 8.6% after matching) At the end of the methods section the authors mention that they used “methods appropriate for the analysis of matched data in estimating the impact of active or passive art and the statistical significance”. However, they continue stating that “Independent t-

	tests were used to assess the impact of active or passive art on wellbeing on participants". Please, clarify. Information on amount of time of involvement and types of activities were transformed in dichotomous variables (engaged or not). This implies an important loss of information. The authors should assess the potential heterogeneity of results when specifying their engagement variables differently (accounting for example for different thresholds in terms of amount of time). Results Many of units were discarded in the PSM analyses. The authors should mention this and provide some interpretation on why units were dropped. The authors could compare descriptive statistics on the confounders of unmatched and matched units. Please, report also the value of the t statistics in Table 4. Discussion An important limitation of using a retrospective cross-sectional design is that it is not possible to match individuals on wellbeing before art engagement was measured. This implies that issues of reverse causality (i.e., that the estimated associations are due to the fact that better off individuals are more likely to engage in art activities and not the other way around) cannot be ruled out. This limitation should be mentioned more explicitly. Strobe statement Point 12e. The authors state that there is no need to implement sensitivity analysis in PSM because they included all known confounders. This statement is too strong.

VERSION 1 – AUTHOR RESPONSE

Reviewer 1: Dr. Emma Wood, School of Events, Tourism and Hospitality Management, Leeds Beckett
Note: Tracked changes from Dr. Wood's suggestions are in orange

No.	Reviewer Feedback	Author's Response	Revision Information
1.1	A well-executed study that adds to the growing evidence base in this area	A heartfelt thanks to Dr. Wood for your encouraging words and feedback for your study. We are thrilled to be able to contribute to this growing area.	N.A.
1.2	Very well written in an accessible style but includes a few typos. A further proofreading needed	The manuscript has been proofread and identified typos amended.	Throughout Manuscript
1.3	Consistency needed in the abstract and earlier sections in how the sampled population are referred to. ie not 'elders' or 'elderly' or 'aged' or 'older' - better to use 'people over 50 years of age' as other terms subjective and relative.	The description of participants has been standardized in the manuscript based on your feedback.	Page 2, 3

1.4	Demographic data might usefully have included household composition as living alone likely to be significant factor in subjective wellbeing. If this is identified from the interpersonal support scale it would be good to have this explained	The demographic information has been updated to reflect the self-reported living arrangement of the participants. 6.7% of the participants reported living alone, but this was not controlled for in the analyses, and will be reported as a limitation of the study.	Page 9, 15
1.5	The discussion of findings and conclusions needs to include a commentary on the direction of the causal relationship for several of the significant results. For example, a sense of belonging may be a precursor to art activity rather than a result of it. The implications for this would be that an enabler of arts participation is to first facilitate belonging. The same holds true for several of the other wellbeing, QoS measures.	Thank you for highlighting this limitation and providing a detailed example. We have added a commentary on potential reverse causality and its implication in the limitation section.	Page 15

Reviewer 2: Dr. Helen Chatterjee, University College London, UK

Note: Tracked changes from Dr. Chatterjee's suggestions are in green

No.	Reviewer Feedback	Author's Response	Revision Information
2.1	This as an interesting population sampling approach using observational and PSM analysis to address the value of creative ageing in Singapore. There are many positive aspects to the study - not least the fact that this is the first comprehensive study based on an Asian population sample, but the paper would benefit from drawing on a wider range of literature from the extensive evidence base. Overall the manuscript is well written, clear and informative; the study adds value to the existing evidence base by using self-report measures to correlate with past arts engagement across a small, but significant, population sample.	We greatly appreciate your encouraging and kind words regarding the study, Dr Chatterjee. We are glad to be able to contribute to the existing evidence base for arts engagement and health.	N.A.
2.2	Some of the language is overly subjective, for example in paragraph 1 of the Introduction older adults are described as being 'agonized by depression and dementia' and in places overstates the findings.	The language in the manuscript has been revised.	Throughout Manuscript

2.3	In addition to References 6 and 7 in paragraph 2 of the Introduction I would recommend citing the UK's Creative Health report which brings together over 1000 peer-reviewed and grey literatures on arts and health; this is the most extensive review on the benefits of arts to health and the authors would no doubt benefit from integrating its findings in the present study. (Creative Health: The Arts for Health and Wellbeing. Inquiry Report. 2017. https://www.artshealthandwellbeing.org.uk/appg-inquiry/). The introduction (para 3) suggests that a literature search has been conducted on creative ageing and health, but this overview fails to mention numerous arts and older adult studies - see below to name but a few - the Creative Health report highlights many others: Coulton, S., Clift, S., Skingley, A. and Rodriguez, J. (2015) Community	Thank you for your generous sharing of literature. We have read the Creative Health report; it was indeed valuable and insightful. We have updated the introduction section based on your recommendations and included additional supporting literature.	Page 3, 4
	singing and health in the older population: A randomised controlled trial, British Journal of Psychiatry, 211, 6, 1–6. Thomson, L. J., Lockyer, B., Camic, P. M. & Chatterjee, H. J. (2017). Effects of a museum-based social prescription intervention on quantitative measures of psychological wellbeing in older adults. Perspectives in Public Health. DOI: https://doi.org/10.1177/1757913917737563 Kattenstroth, J-C., Kalisch, T., Holt, S., Tegenthoff, M. & Dinse, H. R. (2013). Six Months of Dance Intervention Enhances Postural Sensorimotor and Cognitive Performance in Elderly without Affecting Cardio-respiratory Functions. Frontiers in Aging Neuroscience, 5 (5).		
2.4	In Methods the authors should define more clearly what they mean by ‘active’ and ‘passive’ engagement for each of the 8 art forms.	Definitions of passive and active engagement were revised, and examples provided for better clarity.	Page 5, 6

2.5	The Results compared findings with a control group but this is not mentioned in the Study Design; the Methods should include an explanation of the control group and how these data were sourced.	The control group was introduced in the methods section under the statistical analyses sub-header (specifically in the writeup for propensity score matching), however it was introduced as the 'comparison group'. This has since been revised to 'control group' throughout the manuscript for consistency and to avoid confusion.	Page 7
2.6	Further information is needed about how participant health status was assessed.	Details of participant health status assessment has been included in the manuscript.	Page 6
2.7	The Discussion presents a useful synthesis of some of the mechanisms underpinning the value of arts engagement to health which is very useful, but it would be valuable to highlight the range of variation in the various art forms involved and contrast this with previous studies.	Thank you for your suggestion, a writeup of this point has been included in the manuscript.	Page 14
2.8	There should also be recognition of the limitations of the study, including the use of incentives to encourage participation and the accuracy of memory recall for activities undertaken at least 3-6 months prior to survey data collection.	Thank you for recognizing these limitations, they have been included in the manuscript. During the study, a cash voucher worth SGD$20 was presented to the participants upon survey completion. This was a small token of appreciation to	Page 15
		thank participants for their time, and researchers were trained not to use the token of appreciation as an incentive to recruit participants. Participants in the face-to-face interviews appeared to be interested in sharing their opinions with the researchers, and some perceived the cash voucher as a bonus remuneration for their time.	

REVIEWER 3: DR. BRUNO ARPINO, UNIVERSITY OF FLORENCE, ITALY
Note: Tracked changes from Dr. Arpino's suggestions are in blue.

No.	Reviewer Feedback	Author's Response	Revision Information
3.1	This is an interesting study that about the effect of active and passive engagement on several measures of wellbeing of older adults in Singapore.	Our sincere gratitude to Dr. Arpino for your interest and your detailed feedback for the study.	N.A.

3.2	Abstract: All reported associations between art engagement and wellbeing outcome measures are positive but t statistics are negative. Please, double check.	Thank you for pointing out this error. student ‘t’ test value is |t| value i.e., absolute magnitude. We have now removed the negative sign.	Page 2, 10, 11
3.3	Methods: Please clarify whether or not all individuals age 50+ were interviewed in each randomly selected household.	In each household, only one persons above the age of 50 was interviewed in the randomly selected household before moving on to the next household. This information has been included into the manuscript for better clarity.	Page 5
3.4	Methods: One important confounding factor is participation in activities different than art. The authors collected data on physical activity but it is not clear whether authors collected data on social activities, such as volunteering and participating in a (non art) club or association. If information on other activities was not collected, the authors should justify this choice and acknowledge in the paper limitation that such an important confounding factor may bias the estimates. For example, individuals involved in art activities may also be involved in volunteering and the effect on wellbeing may simply be the effect of engagement in general and not of art engagement per se.	As the scope of this survey was limited to arts engagement and physical activity, data on other social activities were not recorded. We would definitely include this for future studies, but for this manuscript, it will be reported as a limitation.	Page 15
3.5	Methods: Related to the previous comment, it is not clear actually	To clarify, involvement in other activities were not assessed in this study. Our	Page 6

	whether involvement in other activities is part of some outcome measures. On page 6 the authors mention “daily activities” and “social functioning” as components of the first two outcome measures, respectively. Thus, it seems that a potential confounding factors is in fact considered as part of the outcome. The authors should clarify the role of these other activities and social functioning (should they be considered as confounders or as outcomes?) and also how the different components of the outcome measures are defined and measured.	apologies that this was not clearly mentioned in our manuscript. ‘Daily activities’ was an assessment of the participant’s perceived quality of their daily living activities, one of the quality of life domain that was assessed by the WHOQOL-8. A sample question for ‘daily activities’ would include “How satisfied are you with your ability to perform your daily living activities?”. In addition, ‘Social functioning’ in this case was an assessment of the participant’s perceived social functioning in relation to their health status, as assessed by the SF-20. A sample question for this would be “How much of the time has your health limited your social activities (such as visiting friends or close relatives)?”. The above items were assessments of social wellbeing and quality of life, and were not assessments of involvement in other activities, hence they were outcome measures in this study instead of confounders. Thank you for highlighting this, we have revised this section to avoid confusion for our readers.	
3.6	The authors should provide much more detail on the implementation of the PSM. What is the estimand of interest (ATE or ATT)? Did the author use matching with or without replacement? What balance diagnostic was considered? I guess it was the standardized mean difference between treatment and control group but the authors should be more explicit on this point.	Apologies for not having clarity in this section. Our estimand of interest was average treatment effect on the treated (ATT). Matching was done without replacement. The standardized mean difference and overall median bias were used as balance diagnostic. Now, we have added more sentences to clarify it.	Page 7, 8

3.7	They should also check table 3: before matching it seems that they show the unstandardized mean difference. Moreover, the reported values of standardized difference do not seem to match with those reported in the text (e.g., median biases of 1.4% and 8.6% after matching)	Thank you for the clarification. Yes, before matching, we presented unstandardized mean difference and now it is labelled as well. We can consider the standardized difference as main balance statistics. However, median bias (before and after) is a crude measure which is usually considered below 10% and easier for layman perspective. In any PSM analysis, bias after matching cannot be zero due to unobserved confounding in observational studies. Successful matching is indicated when the absolute	Page 11

		standardised differences of means is less than 0.25. you can see in the table 3, most of our variables are below this threshold.	
3.8	Methods: At the end of the methods section the authors mention that they used “methods appropriate for the analysis of matched data in estimating the impact of active or passive art and the statistical significance”. However, they continue stating that “Independent t-tests were used to assess the impact of active or passive art on wellbeing on participants”. Please, clarify.	In matched samples, our aim was to compare means between the active/passive and control groups as all our outcome variables were continuous in nature. Therefore, we simply used t-test to compare those groups. This useful and simple analysis served our purpose. Similar analysis has also been done previously. Please see below for some references; https://bmjopen.bmj.com/content/9/1/e022544#ref-27 https://bmcanesthesiol.biomedcentral.com/articles/10.1186/s12871-017-0398-z	Page 7, 8
3.9	Information on amount of time of involvement and types of activities were transformed in dichotomous variables (engaged or not). This implies an important loss of information. The authors should assess the potential heterogeneity of results when specifying their engagement variables differently (accounting for example for different thresholds in terms of amount of time).	Our aim was to compare the outcome variables between who did engage in some form of activities (active or passive) and who did not. Therefore, for this purpose, we have dichotomised activity variables. Effect of amount of time spent on some activity can easily be seen in the form of higher/lower mean score of outcome variables and we compared it. For example, higher activity may produce higher quality of life and wellbeing scores, it can easily be observed between these predefined groups.	N.A.

3.10	Results: Many of units were discarded in the PSM analyses. The authors should mention this and provide some interpretation on why units were dropped. The authors could compare descriptive statistics on the confounders of unmatched and matched units.	Thank you for this very useful suggestion. We have now added appropriate text explaining that PSM analysis has clearly distinguished matched samples with the unmatched samples by comparing descriptive statistics as suggested. We have added a supplementary table showing comparison of confounders between matched and unmatched samples in both active and passive groups.	Page 11
3.11	Results: Please, report also the value of the t statistics in Table 4.	The t-statistics value has been updated in Table 4 as advised.	Page 13
3.12	Discussion: An important limitation of using a retrospective cross-sectional design is that it is not possible to match individuals on wellbeing before art engagement was measured. This implies that issues of reverse causality (i.e., that the estimated associations are due to the fact that better off	Thank you for highlighting this limitation and providing a detailed example. We have revised this point to be mentioned more explicitly in the manuscript.	Page 15
	individuals are more likely to engage in art activities and not the other way around) cannot be ruled out. This limitation should be mentioned more explicitly.		
3.13	Strobe Statement: Point 12e. The authors state that there is no need to implement sensitivity analysis in PSM because they included all known confounders. This statement is too strong.	We checked the model sensitivity for robust estimation and no other additional sensitivity analysis was performed due limited scope of any useful sensitivity analysis in this study. We apologize for that strong statement. We have modified the statement accordingly.	Page 8, 10; Revised Strobe Statement Point 12e

VERSION 2 – REVIEW

REVIEWER	Bruno Arpino University of Florence
REVIEW RETURNED	06-Jul-2019

GENERAL COMMENTS	I would like to thank the authors for having carefully considered all my comments. The authors have satisfactorily replied to most of my comments. However, I have some remaining relatively minor comments. On page 7 – lines 54/55 the authors wrote: “Successful matching was indicated when the absolute standardized mean difference before and after matching was less than 0.25.” However, the word

	"before" should be removed because the balance before matching cannot change due to the matching method used. On page 11 – lines 12/15 the authors wrote: "…unmatched units were dropped as the propensity score model did not find any appropriate control with respect to the case." The authors used a nearest neighbor matching without replacement but they did not specify whether a caliper was imposed or not. The authors' statement "did not find any appropriate control with respect to the case" makes me thinking that a caliper was imposed. But if a caliper was not imposed, treated units can be remain unmatched because of the imposition of the "without replacement" or "common support" options. Please clarify. In my previous report I asked for a clarification about the (un)standardized differences presented in Table 3. The authors clarified that they provide the unstandardized differences before matching and the standardized ones after matching. However, from table 3 it seems that in both cases the reported differences are not standardized, i.e. they are simple differences between means or proportions between treated and controls. Additionally, there is no reason to provide the unstandardized difference before matching. To help comparison and assess how much matching removed initial imbalance usually the standardized (absolute) differences are reported before and after matching. These are automatically calculated by the software the authors used.

VERSION 2 – AUTHOR RESPONSE

1. REVIEWER COMMENTS: DR. BRUNO ARPINO, UNIVERSITY OF FLORENCE, ITALY

No.	Reviewer Feedback	Author's Response	Revision Information
2.1	On page 7 – lines 54/55 the authors wrote: "Successful matching was indicated when the absolute standardized mean difference before and after matching was less than 0.25." However, the word "before" should be removed because the balance before matching cannot change due to the matching method used.	Thank you for pointing out this typo. Now, we have removed this word. It can now be read as; "Successful matching was indicated when the absolute standardized mean difference after matching was less than 0.25.[41]"	Page 7
2.2	On page 11 – lines 12/15 the authors wrote: "…unmatched units were dropped as the propensity score model did not find any appropriate control with respect to the case." The authors used a nearest neighbor matching without replacement but they did not specify whether a caliper was imposed or not. The authors'	Thank you for your comment and asking for clarification on this point. We'd like to highlight that we have already clearly mentioned these points in the statistical analysis section and read as given below;	Please refer to page 7

	statement “did not find any appropriate control with respect to the case” makes me thinking that a caliper was imposed. But if a caliper was not imposed, treated units can be remain unmatched because of the imposition of the “without replacement” or “common support” options. Please clarify.	“The active and passive groups were matched with the control group on the logit of the propensity score, using calipers of width equal to 0.2 of the standard deviation.[39] A nearest neighbor matching technique without replacement was carried out using the Stata ‘psmatch2’ module.[40]”	
2.3	In my previous report I asked for a clarification about the (un)standardized differences presented in Table 3. The authors clarified that they provide the unstandardized differences before matching and the standardized ones after matching. However, from table 3 it seems that in both cases the reported differences are not standardized, i.e. they are simple differences between means or proportions between treated and controls. Additionally, there is no reason to provide the unstandardized difference before matching. To help comparison and assess how much matching removed initial imbalance usually the standardized (absolute) differences are reported before and after matching. These are automatically calculated by the software the authors used.	Thank you for your point. We’d like to highlight that Stata’s ‘psmatch2’ command provides standardised means before and after matching for each matching variable (for both treated and control groups) in the model. Therefore, the difference between these standardised means provides the standardised differences. Similar approach has been applied by several different authors using Stata or any other statistical software (please see, Girard & Farkas 2019; Mason et al 2019; Bheeshma et al 2013). Similarly, before standardised before matching are also commonly reported in propensity score matching articles (again please see, Girard & Farkas 2019; Mason et al 2019; Bheeshma et al 2013). It is well documented and recommended in the literature of reporting and guidelines for propensity score analysis (see Yao et al 2017 for more details). Therefore, we’d like to report both matched and unmatched differences for better clarity and readers point of view. References: Girard L, Farkas C. Breastfeeding and behavioural problems: Propensity score matching with a national cohort of infants in Chile. *BMJ Open* 2019;**9:**e025058. doi: 10.1136/bmjopen-2018-025058 Mason C, Sabariego C, Thắng ĐM *et al*. Can propensity score matching be	N.A.

		applied to cross-sectional data to evaluate Community-Based Rehabilitation? Results of a survey implementing the WHO's Community-Based Rehabilitation indicators in Vietnam. *BMJ Open* 2019;**9:**e022544. doi: 10.1136/bmjopen-2018-022544 Bheeshma R, Ruth C, Austin PC et al. The relation between total joint arthroplasty and risk for serious cardiovascular events in patients with moderate-severe osteoarthritis: propensity score matched landmark analysis *BMJ* 2013;347:f6187. doi: 10.1136/bmj.f6187 Yao XI, Wang X, Speicher PJ et al. Reporting and Guidelines in Propensity Score Analysis: A Systematic Review of Cancer and Cancer Surgical Studies. *J Natl Cancer Inst*. 2017;109(8):djw323. doi:10.1093/jnci/djw323	

VERSION 3 – REVIEW

REVIEWER	Bruno Arpino University of Florence
REVIEW RETURNED	13-Sep-2019

GENERAL COMMENTS	I thank the authors for having considered my previous comment. However, in Table 3 the reported differences are still the simple unstandardized differences between the mean of each variable in the treatment and control group. The authors indicate this difference as "unstandardized" in the before matching column and as "standardized" in the after matching column. The authors should replace these statistics with their standardized versions and change the label in the before matching column accordingly.

VERSION 3 – AUTHOR RESPONSE

2. REVIEWER COMMENTS: DR. BRUNO ARPINO, UNIVERSITY OF FLORENCE, ITALY

No.	Feedback	Author's Response	Revision Information
1.1	I thank the authors for having considered my previous comment. However, in Table 3 the reported differences are still the simple unstandardized differences between the mean of each variable in the treatment and control group. The authors indicate this difference as "unstandardized" in the before matching column and as "standardized" in the after matching column. The authors should replace these statistics with their standardized versions and change the label in the before matching column accordingly.	We apologise for this issue and thank you for your consistent support in reviewing our manuscript. We have now added standardised differences as suggested in the table 3 for both unmatched and matched groups with correct labelling. We used the approach suggested by Yang and Dalton for calculation of standardized differences. Yang DS, Dalton JE. A Unified Approach to Measuring the Effect Size Between Two Groups Using SAS. SAS Global Forum 2012. Paper 335 https://www.lerner.ccf.org/qhs/software/lib/stddiff.pdf	Page 12