Allison et al. Lipids in Health and Disease (2016) 15:77
DOI 10.1186/s12944-016-0242-3

Lipids in Health and Disease

COMMENTARY Open Access

Incorrect statistical method in parallel-groups RCT led to unsubstantiated conclusions

David B. Allison[1,4*], Lisa H. Antoine[2] and Brandon J. George[3]

Abstract

The article by Aiso et al. titled "Compared with the intake of commercial vegetable juice, the intake of fresh fruit and komatsuna (*Brassica rapa* L. var *perviridis*) juice mixture reduces serum cholesterol in middle-aged men: a randomized controlled pilot study" does not meet the expected standards of *Lipids in Health and Disease*. Although the article concludes that there are some significant benefits to their komatsuna juice mixture, these claims are not supported by the statistical analyses used. An incorrect procedure was used to compare the differences in two treatment groups over time, and a large number of outcomes were tested without correction; both issues are known to produce high rates of false positives, making the conclusions of the study unjustified. The study also fails to follow published journal standards regarding clinical trial registration and reporting.

Keywords: Statistical analyses, Vegetable, Fruit, Cholesterol, Nutrition

Background

The conduct of rigorous randomized controlled trials (RCTs) is essential for progress in nutrition-related research [1]. In particular, rigorous tests of the causal effects of fruit and vegetable consumption on aspects of health would be valuable [2]. We therefore read with interest the paper by Aiso et al. [3] reporting results of an RCT of the effects of consumption of a commercial vegetable juice to that of the intake of fresh fruit and komatsuna (*Brassica rapa L. var. perviridis*) juice on serum cholesterol in men. Unfortunately, upon reading it became clear that incorrect statistical analyses were used, that the conclusions drawn in the paper are not supported by the analyses reported, and that there is insufficient adherence to RCT reporting guidelines [4], making it further difficult to determine the appropriateness of the analyses and the extent to which they adhere to original analytic plans.

What the authors conclude

The authors conclude "Compared with the intake of commercial vegetable juice, the intake of fresh fruit and *B. rapa* juice is highly effective in reducing serum cholesterol." As we will show below, this conclusion is not supported by the data and analyses presented.

Why the analysis is incorrect

The stated goal of this study was to compare the effects of the two types of juices on anthropometric data, blood constituents, and dietary intake. To do so, the authors performed paired tests (baseline versus after 4 weeks) within each treatment group, and declared a significant difference between the juices when one juice's test came up significant and the other juice's test did not. This analysis strategy is frequently used in published literature, but is not statistically valid and can result in a type-1 error rate as high as 50% in trials with two groups [5]. As Allison et al. [6] wrote, given a parallel-groups RCT with measures of a continuous outcome at baseline and at endpoint; there are at least four legitimate ways to formally test the difference between two groups: (a) ignore the baseline data and analyze the endpoint data only with a simple independent samples t-test; (b) use a repeated measures ANOVA with one between-groups

* Correspondence: dallison@uab.edu
[1]Nutrition Obesity Research Center and Department of Biostatistics, University of Alabama at Birmingham, Ryals Public Health Building, Room 140J, Birmingham, AL 35294, USA
[4]School of Public Health, University of Alabama at Birmingham, Ryals Public Health Building, Room 140J, Birmingham, AL 35294, USA
Full list of author information is available at the end of the article

factor (treatment assignment) and one within-groups factor (time) and test the group-by-time interaction ($Y_{ij} = \beta_0 + \beta_1 Treatment_i + \beta_2 Time_j + \beta_3 Treatment_i Time_j + e_{ij}$ *for* $i = 1, \ldots, N,\ j = 0, 1,$ *and* $\{e_{ij}\}$ *has a multivariate normal distribution*) [7, 8]; (c) analyze change scores (i.e., endpoint measurement minus baseline measurement) with a simple independent samples t-test; or (d) analyze the final outcome as an ANCOVA with one between-groups factor (treatment assignment) and one covariate (baseline scores) [9]. More details on these methods can be found in many classic experimental design books [7, 9, 10] and tutorial papers [6, 8, 10, 11]. Of note, method (d) (ANCOVA) is typically more powerful than method (c) (t-test on change scores) as it uses the observed pre-post correlation to more efficiently reduce the residual variance [11–13].

Why the conclusions of the paper are not supported

Because a proper test *between groups* was not reported, we emailed the corresponding author of the paper, explained the statistical concern, and requested the standard deviation for the change in LDL-cholesterol and change in total cholesterol in each group or that they make the raw data available thereby allowing us to calculate the values ourselves. Unfortunately, we received no reply to our request. The ICMJE guidelines (http://www.icmje.org/icmje-recommendations.pdf) state "authors have a responsibility to respond appropriately and cooperate with any requests from the journal for data or additional information should questions about the paper arise after publication." Given this, we suggest that Aiso et al. make the raw data from this trial available so that others may verify the results.

Although appropriate between-groups tests of the effects of treatment assignment on the key outcome variables were not reported, it seems unlikely that many of such tests could be significant. For total and LDL cholesterol on which Aiso et al's conclusion claim is based, Aiso et al. do report the means and standard deviations for each variable within each of the treatment and control groups both at baseline and at endpoint. Using this information, we can implement choice *a* above[1]. If we do this for total cholesterol, the two-tailed *p*-value is 0.9480 ($t = 0.0663$; df = 14). If we do this for LDL cholesterol, the two-tailed *p*-value is 0.5525 ($t = 0.6087$; df = 14). In neither case is the result even close to significant, meaning that by this legitimate test, the appropriate conclusion would have been that there was no compelling evidence of a treatment effect.

Admittedly, the t-test only on endpoints is a relatively low power test; choice *c* above (a t-test on change scores) will usually be more powerful. Although it is clear that such a t-test would not be significant for LDL cholesterol (the groups had identical 9 mg/dl reductions), it is conceivable that the difference between the two groups in change of total cholesterol is statistically significant but we lack necessary information (such as the standard deviation of the change score) to conduct such a test. If Aiso et al. can show a statistically significant between-groups difference in the outcome variable, then their conclusion would be supported, but at present it is unsupported.

There is a concern regarding Aiso et al.'s reporting of *p*-values from 58 variables per treatment group (116 tests overall). Such a high number of tests would strongly suggest the use of a multiple testing correction to control the type-1 error rate [14] as one may expect approximately 5.8 significant findings to occur by chance alone if one tests 116 independent tests with a significance level of 0.05 and all the null hypotheses are true (i.e., there is really nothing to find). The smallest reported *p*-value was 0.012, far larger than what would be needed for significance under a Bonferroni (0.000431) or Sidak [15] (0.000442) correction. Although correlation between the 58 variables may reduce the extent of Type I error inflation and methods exist for correcting multiple correlated outcomes [16], those methods were not used in this article and without knowing the correlation between each variable it is impossible to quantify the extent of the inflation. Taken as a whole, it is plausible that many of the *p*-values reported as significant represent type-1 errors.

Lack of trial registration

Articles published in *Lipids in Health and Disease* require adherence to BioMed Central's editorial policies, http://www.lipidworld.com/about. BioMed Central follows the International Committee of Medical Journal Editors (ICMJE) guidelines, which necessitate clinical trials registration for RCT reports submitted to its journals. ICMJE defines a clinical trial as, "any research study that prospectively assigns human participants or groups of humans to one or more health-related interventions to evaluate the effects on health outcomes" [17]. ICMJE recommends that authors include the trial registration number in the abstract of the manuscript. This journal article does not include the clinical trials registration number. We emailed the authors to inquire about public clinical trials registry for this article, but received no response. Given the above, we believe that the authors should provide documentation of clinical trial registration.

Conclusions

Clinicians, scientists, regulators, and the general public require and have a right to expect scientific evidence based on valid procedures [18] and free from spin [19] on which they can base decisions. The Committee on Publication Ethics [20] states that "Journal editors should consider

retracting a publication if…they have clear evidence that the findings are unreliable [including]… as a result of … miscalculation or experimental error." We believe that the conclusions of Aiso et al. [3] are unreliable as a result of using an incorrect statistical procedure.

Endnotes

[1]We conducted our calculations with the free public software at this site http://www.graphpad.com/quickcalcs/ttest2/ so that anyone could reproduce our calculations.

Abbreviation

RCT: Randomized controlled trial.

Competing interests

The authors report no financial connection to the content of the paper discussed. David B. Allison and/or his institution have accepted funds from food companies, but not ones who, to his knowledge, market products discussed in this research.

Authors' contributions

David B. Allison conceived the paper. All three authors drafted sections of the manuscript and edited the entire paper. All authors read and approved the final manuscript.

Author details

[1]Nutrition Obesity Research Center and Department of Biostatistics, University of Alabama at Birmingham, Ryals Public Health Building, Room 140J, Birmingham, AL 35294, USA. [2]School of Engineering, University of Alabama at Birmingham, Birmingham, AL 35294, USA. [3]Department of Biostatistics, University of Alabama at Birmingham, Birmingham, AL 35294, USA. [4]School of Public Health, University of Alabama at Birmingham, Ryals Public Health Building, Room 140J, Birmingham, AL 35294, USA.

Authors' information

All authors are affiliated with the University of Alabama at Birmingham. David B. Allison is Associate Dean for Science in the School of Public Health, Distinguished Professor of Biostatistics, and Director of the NIH-funded Nutrition Obesity Research Center. Lisa H. Antoine is a doctoral student in Interdisciplinary Engineering. Brandon J. George is a statistician and holds a PhD in Biostatistics.

Acknowledgements

Supported in part by NIH grants P30DK056336, R25DK099080, and R25HL124208. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or any other organization.

Response to "Incorrect statistical method in parallel-groups RCT led to unsubstantiated conclusions"

Izumi Aiso[1†], Hiroko Inoue[2†], Yukiko Seiyama[3], Toshiko Kuwano[3*]

[1]Shizuoka Prefectural Chubu Public Health Center
[2]Department of Nutrition and Health Sciences, Faculty of Food and Nutritional Sciences, Toyo University
[3]Department of Food and Nutritional Sciences and Environmental Health Sciences, Graduate School of Integrated Pharmaceutical and Nutritional Sciences, University of Shizuoka, 52-1 Yada, Suruga-ku, Shizuoka 422-8526, Japan

[†]I. Aiso and H. Inoue contributed equally to this work.
[*]Corresponding author

Introduction

Here we respond to the commentary on our article by Allison et al. As an overview of the issues they raise, we have selected the following statement from their commentary.

"Although the article concludes that there are some significant benefits to their komatsuna juice mixture, these claims are not supported by the statistical analyses used. An incorrect procedure was used to compare the differences in two treatment groups over time such that no direct between-group comparison was done, and a large number of outcomes were tested without correction."

At various points their commentary, they raise questions about the following aspects of our study: 1) statistical analysis, 2) dietary survey, and 3) clinical trial registration. Below, we respond to each of these questions in turn.

Statistical analysis

Before we turn to the detailed statistical issues that Allison et al. raise in their commentary, we would first like

to point out that, when our article was reviewed by the journal's referees, both referees specifically stated that our statistical analysis did not need further verification. Referee 1 wrote in his/her review:

"**Statistical review:** No, the manuscript does not need to be seen by a statistician."

And Referee 2 wrote in his/her review:

"**Statistical review:** No, the manuscript does not need to be seen by a statistician."

As the referees thought that a statistical review was unnecessary, we felt confident that we had used the appropriate statistical methods. Therefore, we were surprised that our methods were criticized by Allison et al. However, we were also very interested in their interpretation, so we followed their suggested method for analyzing our data.

In their commentary, Allison et al. state:

"The stated goal of this study was to compare the effects of the two types of juices on anthropometric data, blood constituents, and dietary intake. To do so, the authors performed paired tests (baseline versus after 4 weeks) within each treatment group, and declared a significant difference between the juices when one juice's test came up statistically significant (defined by the authors as $p<0.05$) and the other juice's test did not. This analysis strategy of comparing nominal significance of within group changes (done with either paired-sample parametric or non-parametric tests) is frequently used in published literature, but is not statistically valid and can result in a false positive rate as high as 50% in trials with two groups of equal size."

and

"there are several legitimate ways to formally test the difference between two groups: (a), (b), (c)."

[…]

"Method (c) (ANCOVA) is typically more powerful than method (b) (t-test on change scores). As it uses the observed pre-post correlation to more efficiently reduce the residual variance."

We thank Allison et al. for pointing out this new statistical approach. Like the other researchers in the field, we were not aware of this more advanced method when we wrote our paper. However, we were pleased to be able to apply it to the analysis of our data following the suggestion of Allison et al. We return to that analysis below, but first we would like to clarify some issues related to the Wilcoxon signed rank test that we used in our study.

The goal of this study was to examine changes in various parameters in the intervention group and the control group before and after their respective juice interventions. In analyzing our data, we performed both a paired t-test and a Wilcoxon signed rank test. Both tests showed that the concentration of total cholesterol and LDL-cholesterol in the intervention group were significantly lower after 4 weeks compared with the baseline values. However, we chose to report the results of the Wilcoxon signed rank test, as that test is more appropriate for small sample sizes that do not have normal distributions. In retrospect, we now think that it would have been clearer to include the reason that we selected the Wilcoxon test in our article. We apologize for any misunderstanding that this omission may have caused.

As Allison et al. raised questions about our statistical methods, we first rechecked the results of our original statistical analysis. After that, we applied the statistical analysis that they have suggested.

a. Re-application of Wilcoxon signed rank test. When we repeated the Wilcoxon signed rank test to ensure that the results of our original test were accurate, we found that the results were the same as those we had originally published.

b. Application of analysis of covariance (ANCOVA). In their commentary, Allison et al. state:

"there are several legitimate ways to formally test the difference between two groups: (a), (b), (c)."

[…]

"Method (c) (ANCOVA) is typically more powerful than method (b) (t-test on change scores). As it uses the observed pre-post correlation to more efficiently reduce the residual variance."

Following this suggestion, we re-examined our data using ANCOVA. Data were analyzed using SPSS for Windows version 15.0 J computer software. We conducted the ANCOVA test by adjusting for levels of age, BMI, and each variable. The results did not show a significant difference for the parameters in either group. We speculate that this may be due to the small number of subjects participating in the study. In the light of this new analysis, we think that the conclusions stated in our original article should be moderated. We return to this point in our conclusion below.

Dietary survey

In their commentary, Allison et al. state:

"There is also a concern regarding Aiso et al.'s reporting of *p*-values from 58 variables per treatment group (116 tests overall)."

We would like to clarify the nature of the dietary survey used in our study. We used the "brief-type self-administered diet history questionnaire" (BDHQ), which is a standard dietary history survey instrument used in Japan [21, 22]. The BDHQ is used to calculate intake values such as energy and nutrients based on information about 58 types of food and drink. The questionnaires are batch processed at the Diet History Questionnaire Support Center. Because the Center calculates nutrient

values from the intake frequency of the 58 food and drink items and then sends those values to us, we cannot individually manipulate the food and drink items as variables afterwards by ourselves. We can only perform our analysis based on the data categories provided by the Center, and therefore we cannot perform a more detailed statistical analysis on the 58 items.

Clinical trial registration

In their commentary, Allison et al. state:

"The study also fails to follow published reporting guidelines for this journal and the scientific community overall regarding clinical trial registration and reporting."

We have registered our study with the UMIN-CTR Clinical Trial database. The information can be found at: https://upload.umin.ac.jp/cgi-open-bin/ctr/ctr.cgi?function=brows&action=brows&recptno=R000022765&type=summary&language=E.

Conclusion

The statistical test that we used in our article was considered appropriate by both of the journal's reviewers. We have rechecked the results of that test, and have obtained the same result. In addition, we have carried out the ANCOVA test suggested by Allison et al. ANCOVA did not show a significant difference between the intervention group and the control group. We thank Allison et al. for providing us with this insight. In future studies we plan to increase the number of subjects and reinvestigate the effect. We will also consider conducting a crossover study similar to that of Lee et al. in the future [23].

Based on the insights that we have gained from the suggestions of Allison et al., we think that the conclusion originally drawn in our article should be moderated to be stated as follows:

Compared with the intake of commercial vegetable juice, the intake of fresh fruit and *B. rapa* juice may be effective in reducing serum cholesterol.

We would like to once again thank Dr. Allison and colleagues for their thoughts on our article.

Authors' information

Izumi Aiso is a pharmaceutical chemist, Hiroko Inoue is a registered dietitian, Ph.D., Yukiko Seiyama is a registered dietitian, and Toshiko Kuwano is a registered dietitian, Ph.D.

Received: 3 December 2014 Accepted: 21 March 2016
Published online: 15 April 2016

References

1. Casazza K, Allison DB. Stagnation in the clinical, community, and Public Health Domain of obesity: the need for probative research. Clin Obes. 2012;2(3-4):83–5.
2. Kaiser KA, Brown AW, Bohan Brown MM, Shikany JM, Mattes RD, Allison DB. Increased fruit and vegetable intake has no discernible effect on weight loss: a systematic review and meta-analysis. Am J Clin Nutr. 2014;100(2):567–76.
3. Aiso I, Inoue H, Seiyama Y, Kuwano T. Compared with the intake of commercial vegetable juice, the intake of fresh fruit and komatsuna (Brassica rapa L. var. perviridis) juice mixture reduces serum cholesterol in middle-aged men: a randomized controlled pilot study. Lipids Health Dis. 2014;13:102.
4. De Angelis CD, Drazen JM, Frizelle FA, Haug C, Hoey J, Horton R, et al. Is this clinical trial fully registered? A statement from the International Committee of Medical Journal Editors. Lancet. 2005;365(9474):1827–9.
5. Bland JM, Altman DG. Comparisons against baseline within randomized groups are often used and can be highly misleading. Trials. 2011;12(264):1–7.
6. Allison DB, Gorman BS, Primavera LH. The most common questions asked of statistical consultants: our favorite responses and recommended readings. Genet Soc Gen Psychol Monogr. 1993;119:153–85.
7. Winer BJ, Brown DR, Michels KM. Statistical Principles in Experimental Design. 3rd ed. New York: McGraw-Hill; 1991.
8. Liu S, Rovine MJ, Molenaar PC. Selecting a linear mixed model for longitudinal data: repeated measures analysis of variance, covariance pattern model, and growth curve approaches. Psychol Methods. 2012;17(1):15–30.
9. Kirk RE. Experimental Design: Procedures for the Behavioral Sciences. 2nd ed. Pacific Grove: Brooks/Cole; 1982.
10. Albert PS. Tutorial in biostatistics: longitudinal data analysis (repeated measures) in clinical trials. Stat Med. 1999;18:1707–32.
11. Huck SW, McLean RA. Using a repeated measures ANOVA to analyze the data from a pretest-posttest design: a potentially confusing task. Psychol Bull. 1975;82(4):511–8.
12. Myers JL, Well AD. Research Design and Statistical Analysis. 1st ed. New York: HarperCollins; 1991.
13. Allison DB. When is it worth measuring a covariate in a randomized clinical trial? J Consult Clin Psychol. 1995;63(3):339–43.
14. Young SS, Bang H, Oktay K. Cereal-induced gender selection? Most likely a multiple testing false positive. Proc Biol Sci. 2009;276(1660):1211–2.
15. Sidak ZK. Rectangular confidence regions for the means of multivariate normal distributions. J Am Stat Assoc. 1967;62(318):626–33.
16. Sankoh AJ, Huque MF, Dubey SD. Some comments of frequently used multiple endpoint adjustment methods in clinical trials. Stat Med. 1997;16:2525–42.
17. ICMJE: Clinical Trial Registration. http://www.icmje.org/recommendations/browse/publishing-and-editorial-issues/clinical-trial-registration.html. [Accessed 30 Mar 2016].
18. McNutt M. Journals unite for reproducibility. Science. 2014;346(6210):679.
19. Boutron I, Altman DG, Hopewell S, Vera-Badillo F, Tannock I, Ravaud P. Impact of Spin in the Abstracts of Articles Reporting Results of Randomized Controlled Trials in the Field of Cancer: The SPIIN Randomized Controlled Trial. J Clin Oncol. 2014 Nov 17. pii: JCO.2014.56.7503.
20. Committee on Publication Ethics. Guidelines for retracting articles. http://publicationethics.org/files/retraction%20guidelines.pdf. [Accessed 30 Mar 2016].
21. Kobayashi S, Murakami K, Sasaki S, et al. Comparison of relative validity of food group intakes estimated by comprehensive and brief-type self-administered diet history questionnaires against 16 d dietary records in Japanese adults. Public Health Nutr. 2011;14:1200–11.
22. Kobayashi S, Honda S, Murakami K, et al. Both comprehensive and brief self-administered diet history questionnaires satisfactorily rank nutrient intakes in Japanese adults. J Epidemiol. 2012;22:151–9.
23. Lee JT, Moore CE, Radcliffe JD. Consumption of calcium-fortified cereal bars to improve dietary calcium intake of healthy women: randomized controlled feasibility study. PLoS One. 2015;10(5):e0125207.