PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Relationships Between Black Tea Consumption and Key Health Indicators in the World: an Ecological Study
AUTHORS	BERESNIAK, ARIEL ; Duru, Gerard; Berger, Genevieve; BREMOND-GIGNAC, Dominique

VERSION 1 - REVIEW

REVIEWER	Andrew Odegaard Research Associate University of Minnesota, USA
REVIEW RETURNED	12-Dec-2011

THE STUDY	As far as I can tell I think the statistical methods are OK. While using PCA in my own research I haven't seen it used like this before, but the description is adequate.
GENERAL COMMENTS	Considering Lipton tea is part of the Unilever's portfolio, and they're providing the tea data, it would be prudent of the author's to note this at the top of their competing interest with a clear statement rather than in an incongruous sentence in the middle. Abstract What about reports on the other key indicators of health as mentioned in the study objective? What is the correlation of black tea consumption with those? The language about direct cause-effect relationship should be removed in the conclusion. Furthermore, so should "obesity" as that wasn't an aim or discussed in the results and seems to conflate diabetes with obesity. As well, ecological research can certainly contribute to the evidence base, but to start truly establishing a "causal" relation, randomized studies should theoretically be performed. Background Last sentence in paragraph- line 26- a reference should be provided on the positive health effects observed. The sentence beginning on line 30 needs reworking so it is clearer for the reader. What does according to longitude mean? There could be multiple interpretations. Overall this whole 2nd paragraph needs attention on the writing. What are the key health indicators being examined? Note- Need general attention to writing/grammar.

	Data sources What are the countries with the highest black tea consumption? Lowest? This would be informative for the reader. How valid are the consumption data? How are they specifically derived? This is necessary information to make any relevant conclusions on the data for the reader. What year(s) are the consumption data from? What year(s) are the health indicator data from? What do these data actually look like on a per country basis and overall? Why is obesity noted in the conclusions of the abstract but is not one of the 5 key indicators? A more thorough description of these aspects of the study is necessary and would seem to greatly improve the interpretation of this manuscript. Methods The summary on data mining and correlation techniques is nice for the reader. By the description of the correlation circle, and then viewing it, it would appear black tea consumption is also correlated with higher infectious diseases, then respiratory diseases but not as greatly with either cancer or cardiovascular disease. Would this insinuate that really what this data represent is the stage of development in the country? I.E. higher levels of infectious and respiratory diseases generally associate stronger with less developed countries. Also, there is attention in the text towards “hazardous” interpretation of the data. This seems to be hyperbolism. Since most readers are probably not aware with the analytic techniques applied it would also be informative if the authors provided interpretation of the other data (outcomes) examined. Even if null. This is likely to be more informative and would seem to improve the quality of the paper. Pg 7, first paragraph- attention to the writing and formation of sentence structure could use attention so what is written makes better sense. The statement starting on line 11 on contribution to the axis is why more thorough data should be provided. What does this mean to the audience when they have no idea what the actual levels of black tea consumption are, and what the levels of the different indicators of health are? Especially concerning would be the presence of China- which consumes black tea at a very high relative level globally and may be “winning the diabetes crown” for overall numbers. Please report the actual data for these other indicators. Let the audience have the data so they can decide how strong/not strong the data are. Discussion

	Thorough discussion, albeit brief, would greatly enhance the interpretation of this paper on the sampling and morbidity data- what are strengths/limits? Some discussion is provided but further on actual questions, for example delineating the type of diabetes and what countries involved may be better representative or non-representative…. The question that is not answered but relevant to the aim is, while the statistical relationship between BT and diabetes is "strong", what is the relationship with the other health indicators? This was not reported. I think many readers may actually think the opposite of the rationale provided for mechanistic pathways starting on line 26 "the relevance and mechanism…". The results were not based upon a-priori work, and in an ecological study with grey data, the only mechanism that may be relevant is what this means on this macro-level, not underlying hypothesized biological mechanisms. There are many things that could explain this simple correlation and plenty may not have anything to do with biological activity of black tea. Ignoring them is misleading to the audience. The writing could be cleaned up a bit in this paragraph. The sentence on pg 9, line 8, referring to the obesity epidemic is cited inappropriately. None of the 4 studies cited address the effectiveness of BT consumption on obesity. This sentence should be removed, or cited appropriately if there is evidence to substantiate the claim. Ironically, the authors follow the sentence noting what a couple of the studies actually investigate- not obesity. For that matter, the whole composition and purpose of this paragraph is nebulous. A hodge-podge of tangential points related to tea and diabetes are summarized. What is the point of this for the reader? The paragraph running from page 9, line 51 through first half of page 10 is more relevant discussion, albeit in a meandering way. If it was cleaned up, and focused on the perils of ecological research and framing the spectrum of human/health/population research it would improve the paper. Discussing causality seems a bit of a stretch and like other areas of the paper seems implanted to almost confuse the reader. The last paragraph nicely summarizes the data technique as well, but again, the going back to causality is immaterial. Overall, I have two major suggestions. One, provide greater data on intake levels in the actual countries, overall, etc, as mentioned above. Two, is to really rework this discussion so it addresses what this study actually aims to do- ecological correlations using a somewhat novel approach. The same pages could easily be filled up discussing what these data actually represent, interpretations, limits, etc and would be greatly improved on a scientific level. In its current format, with no a-priori hypothesis, incomplete data reporting and interpretation, with the constant hits at causality, what the research could mean mechanistically, biologically, etc, with mis-citations, not to mention lack of any individual data it reads closer to a propaganda paper looking to boost tea sales than a scientific manuscript. Furthermore, there are multiple parts of the paper where

	attention to writing is needed. Some of these have been pointed out specifically, and generically.

REVIEWER	Butler, Lesley Colorado State University, Department of Environmental & radiological Health Sciences I have no competing interests.
REVIEW RETURNED	27-Dec-2011

GENERAL COMMENTS	Summary Using ecologic data on amount of black tea purchased and the prevalence of 5 disease groups (diabetes, cancer, infectious disease, cardiovascular disease and respiratory disease), the authors used a two-stage process to identify a statistical inverse relationship between amount of black tea purchased and prevalence of diabetes. There were several major and minor concerns identified with the statistical approach, interpretation and references cited that diminished enthusiasm for the observed findings. These concerns are described below. Major concerns (1) Inappropriate application of multiple correspondence analysis (MCA). This reviewer is not familiar with MCA, so I looked up the reference provided on p.5, line 53 (#13 Briand, 2009). Unfortunately, this reference provides an example of an application of MCA. A reference for the method itself is needed. Using reference #13 to judge the application of MCA in the present analysis led this reviewer to question it's appropriateness. For example in the Briand, et al study, "The MCA modeling method was specifically adapted to reduce the five exposure variables to one aggregated exposure indicator…MCA provides a standardized way to reduce complexity." This is quite different from the use of MCA used in the present analysis, where MCA was used to identify a correlation between one exposure variable and 5 outcome variables. From the documentation provided, MCA does not appear to be a method that is used to pull out an individual exposure-disease relationship, as it was used in the present analyses. (2) Unspecific citations for the two data sources (p.4-5). The market research resource for the black tea intake is not publically available (reference #7). I could not locate information about this data source from the reference information provided. As for the WHO disease prevalence information, this reference was also inadequate (#8). There are several WHO surveys listed on the website provided. Which data was used in the present analysis? It appears that the WHO has data on many regions worldwide, but that this information is not country specific, unlike the black tea information. (3) In the Discussion, the authors go back and forth between providing evidence for green and black tea intake/compounds and disease associations. It appears that the authors incorrectly consider evidence for green tea to be synonymous with evidence for black tea. For example (p.8, lines 43-45), the health benefits of green tea are used to support the author's findings for black tea. The references provided on p. 8, line 54 mix references with findings for green and black tea. On page 9, lines 36-40 describe

	studies with findings (reference 44 and 45) supporting the beneficial health effects of green tea (or green tea catechins), but the authors simply state “tea” without specifying type. On line 33 (p. 9), black tea should be specified for the findings reported for reference #40. Minor concerns (1) In the Background, a rationale is needed for the selection of the 5 disease outcomes chosen. Only a rationale for diabetes is provided, although the authors state that the relationship between black tea and diabetes was not an a priori hypothesis (p.8, second paragraph). (2) In the Methods, references are needed for the “correlation circle” paragraph/method (p.6). (3) In the Results, provide a graph that plots each country/region by amount of black tea intake (x-axis) by the prevalence of diabetes (y-axis). Do this for each outcome. This more common graphical presentation would help the audience grasp what is being shown in the less common correlation circle. (4) The heterogeneity of all of the diseases (especially cancer and infectious diseases) is a limitation and may be responsible for the lack of a statistical relationship with black tea. This issue needs to be mentioned in the Discussion.

VERSION 1 – AUTHOR RESPONSE

Reviewer 1: Andrew Odegaard
Research Associate, University of Minnesota, USA

See my comments below. As far as I can tell I think the statistical methods are OK. While using PCA in my own research I haven't seen it used like this before, but the description is adequate.

Comment 2: Considering Lipton tea is part of the Unilever’s portfolio, and they’re providing the tea data, it would be prudent of the author’s to note this at the top of their competing interest with a clear statement rather than in an incongruous sentence in the middle.

Response: Tea consumption data were not generated by Unilever, but were derived from a survey conducted by Euromonitor International, an independent international market research agency. Unilever has provided data from the Euromonitor International report after checking overall quality and internal consistency.
The competing interest statement has been completed as follows:

All authors have completed the Unified Competing Interest form : no support from any organisation for the submitted work , no financial relationships with any organisations that might have an interest in the submitted work in the previous three years. Coauthor G.B is employed at Unilever PLC as Chief R&D Officer and has provided access to the Euromonitor International global tea consumption data without any financial agreement or any grant to support this study, which has been carried out in total independence. There are no other relationships or activities that could appear to have influenced the submitted work.

Comment 3: Abstract: What about reports on the other key indicators of health as mentioned in the

study objective? What is the correlation of black tea consumption with those?

Response: The following sentence has been added to the abstract:
The correlation circle confirmed that the "black tea" vector was negatively correlated with the diabetes vector and was not correlated with any of the other four health indicators.

Comment 4: The language about direct cause-effect relationship should be removed in the conclusion. Furthermore, so should “obesity” as that wasn’t an aim or discussed in the results and seems to conflate diabetes with obesity.

Response: The following sentence has been changed in the abstract:
These results are consistent with biological and physiological studies conducted on the potential effect of black tea on diabetes.

Comment 5: As well, ecological research can certainly contribute to the evidence base, but to start truly establishing a “causal” relation, randomized studies should theoretically be performed.

Response: The following sentence has been added to the abstract:
Further epidemiological research and randomized studies are necessary to investigate the causality.

Comment 6: Background - Last sentence in paragraph- line 26- a reference should be provided on the positive health effects observed.

Response: The following references have been provided:

Bahorun T, Luximon-Ramma A, Neergheen-Bhujun VS, et al. The effect of black tea on risk factors of cardiovascular disease in a normal population. Prev Med. 2011;Dec 16.
Wang ZM, Zhou B, Wang YS, et al. Black and green tea consumption and the risk of coronary artery disease: a meta-analysis. Am J Clin Nutr. 2011 2011;93(3):506-515.

Comment 7: The sentence beginning on line 30 needs reworking so it is clearer for the reader. What does according to longitude mean? There could be multiple interpretations. Overall this whole 2nd paragraph needs attention on the writing.

Response: The sentence has been reworded as follows:
Considering the complexity of implementing international prospective studies and the difficulty of conducting meta-analyses on a large number of heterogeneous local studies, potential correlations between BT consumption and epidemiological data around the world could be investigated by deploying a data mining approach using advanced exploratory statistical methods.

Comment 8: What are the key health indicators being examined?

Response: The sentence has been reworded as follows:
The objective of this original research was to investigate potential statistical relationships between BT consumption and the following five key health indicators: respiratory diseases, infectious diseases, cancer, cardiovascular diseases and diabetes.

Comment 9: Note- Need general attention to writing/grammar.

Response: Writing and grammar have been reviewed.

Comment 10: Data sources - What are the countries with the highest black tea consumption? Lowest? This would be informative for the reader.

Response: The following sentence was added
Highest BT consumptions expressed in kg/ year per inhabitant are observed in Ireland (2.1576), UK (1.8137), Turkey (1.6631) and Russia (1.0668). Lowest BT consumptions are observed in South Korea (0.0007), Brazil (0.001) and China (0.0011).

Comment 11: How valid are the consumption data? How are they specifically derived? This is necessary information to make any relevant conclusions on the data for the reader.

Response: Black tea consumption data around the world has been assessed by a survey conducted by an independent agency specialized in market research (Euromonitor International), which appear to be the most robust and consistent data available worldwide. Consumption data are derived from international black tea trading registries, used by black tea importers to adapt international orders to local sales.

The following sentence has been added to the text:
Consumption data are derived from black tea international trading registries, considering that black tea importers use to adapt international orders to local sales.

Comment 12: What year(s) are the consumption data from? What year(s) are the health indicator data from?

Response: Both consumption and epidemiological data are from 2009.

The text has been changed as follows:
...BT consumption data were derived from a specific international trade survey compiling sales data conducted in 2009 by Euromonitor International...

...Using the 2009 dataset, five key health indicators were selected in 50 countries in both men and women for all age groups...

Comment 13: What do these data actually look like on a per country basis and overall?

Response: Consumption and health indicator data per country are now presented in two new figures.

Figure 1: 2009 Black Tea consumption data in kg/year per inhabitant (source: Euromonitor)
Figure 2: 2009 prevalence (per 100,000) of key health indicators (source: WHO)

Comment 14: Why is obesity noted in the conclusions of the abstract but is not one of the 5 key indicators?

Response: Obesity has been deleted from the conclusion of the abstract.
See response to comment 4.

Comment 15: Methods - The summary on data mining and correlation techniques is nice for the reader. By the description of the correlation circle, and then viewing it, it would appear black tea consumption is also correlated with higher infectious diseases, then respiratory diseases but not as greatly with either cancer or cardiovascular disease. Would this insinuate that really what this data represent is the stage of development in the country?

Response: Among the 5 health indicators, only the diabetes vector was correlated with the black tea consumption vector. The infectious disease vector was not correlated with black tea consumption because the 2 vectors were orthogonal in the 3rd dimension (not visible in this figure), as explained in the text:
Of particular interest was the interpretation of the "infectious disease" vector, which seemed to be close to the BT vector in the two-dimensional correlation circle, but was actually represented by a large angle in the third dimension. The infectious disease vector was also closer to the centre of the correlation circle, confirming the poor meaningful correlations and potentially hazardous interpretations.

Comment 16: Also, there is attention in the text towards “hazardous” interpretation of the data. This seems to be hyperbolism. Since most readers are probably not aware with the analytic techniques applied it would also be informative if the authors provided interpretation of the other data (outcomes) examined. Even if null. This is likely to be more informative and would seem to improve the quality of the paper.

Response: As no statistical links were established with the other four health indicators, it is not possible to discuss and interpret the dataset.

The following sentence has been added:
Consequently, among the five health indicators selected, only the diabetes parameter was correlated with BT consumption and can be submitted to discussion and interpretation. No valid interpretations can be derived from the other four health indicators using this dataset.

Comment 17: Pg 7, first paragraph- attention to the writing and formation of sentence structure could use attention so what is written makes better sense.

Response: The paragraph has been rewritten as follows:
The database was composed of 300 fields representing 6 variables (5 health indicators and BT consumption) in 50 countries. Using normative PCA on this multidimensional table, the overall "quality" (percentage of original variance) of the final projection from 6 dimensions (6 variables) was 59% in 2 dimensions and 74% when projected in 3 dimensions. This confirms that the best representation of the dataset should be in 3 dimensions, which can be described by axes entitled F1, F2 and F3. The "BT consumption" variable provided a high contribution to the construction on axis F3 (81%). The angle of the vector "BT consumption" with axis F3 was only 22°, confirming the very high contribution of this variable on axis F3. Forty two of the 50 countries were related to this axis F3. The

8 countries not contributing to F3 were Brazil, China, Venezuela, Morocco, Colombia, Vietnam, Philippines and Israel, suggesting the absence of any correlation between BT consumption and health indicators in these particular countries.

Comment 18: The statement starting on line 11 on contribution to the axis is why more thorough data should be provided. What does this mean to the audience when they have no idea what the actual levels of black tea consumption are, and what the levels of the different indicators of health are? Especially concerning would be the presence of China- which consumes black tea at a very high relative level globally and may be "winning the diabetes crown" for overall numbers.

Response: The addition of the two new figures (Figure 1 and Figure 2) presenting country data should now facilitate the understanding of tea consumption and epidemiological data. Black tea consumption is relatively low in China, as the Chinese population drinks 30 times more green tea (0.036 kg per capita) than black tea (0.0011 kg per capita).

The following sentence was added:
(as the Chinese population drinks 30 times more green tea (0.036 kg per inhabitant) than black tea).

Comment 19: Please report the actual data for these other indicators. Let the audience have the data so they can decide how strong/not strong the data are.

Response: See response to comment 13.

Comment 20: Discussion - Thorough discussion, albeit brief, would greatly enhance the interpretation of this paper on the sampling and morbidity data- what are strengths/limits? Some discussion is provided but further on actual questions, for example delineating the type of diabetes and what countries involved may be better representative or non-representative….

Response: The strengths and limitations of the World Health Survey are discussed in the first paragraph of the discussion section. An additional comment has been provided concerning Singapore, which is the country with the highest diabetes prevalence of the dataset.

The following sentence was added in the text
According to WHS 2009 data, Singapore is the country with the highest diabetes prevalence with 12,876 cases per 100,000 inhabitants.(Figure 2), which is mainly observed in the Chinese community and is probably due to the intense urban lifestyle in Singapore (Ang, 2011)

Comment 21: The question that is not answered but relevant to the aim is, while the statistical relationship between BT and diabetes is "strong", what is the relationship with the other health indicators? This was not reported.

Response: See response to comment 17. No statistical relationship was established with any of the other health indicators.

Comment 22: I think many readers may actually think the opposite of the rationale provided for mechanistic pathways starting on line 26 "the relevance andmechanism…". The results were not

based upon a-priori work, and in an ecological study with grey data, the only mechanism that may be relevant is what this means on this macro-level, not underlying hypothesized biological mechanisms. There are many things that could explain this simple correlation and plenty may not have anything to do with biological activity of black tea.

Response: The following text has been added:
The proposed epidemiological approach considers the population as the unit of analysis rather than an individual and can be presented as an ecological study, which is considered to be inferior to case-control studies in the context of evidence-based medicine. In an ecological study, no information is available about the individual members of the populations compared, whereas in a case-control study, information is reported for each individual. However, ecological studies can be very useful for international comparisons, while case-control studies are exclusively based on local information. Furthermore, when strong correlations have been established, the results of ecological studies can suggest further evidence-based studies, investigating the relevance and mechanism of the statistical relationship. Various study designs have already been used to assess the potential benefits of tea..

Comment 23: The writing could be cleaned up a bit in this paragraph.

Response: The manuscript has been reviewed

Comment 24: The sentence on pg 9, line 8, referring to the obesity epidemic is cited inappropriately. None of the 4 studies cited address the effectiveness of BT consumption on obesity. This sentence should be removed, or cited appropriately if there is evidence to substantiate the claim. Ironically, the authors follow the sentence noting what a couple of the studies actually investigate- not obesity.

Response: The sentence has been revised as follows:
Although many laboratory studies have observed physiological effects of BT on glucose metabolism 30, 31, 36, 37, the underlying mechanisms remain unclear.

Comment 25: The paragraph running from page 9, line 51 through first half of page 10 is more relevant discussion, albeit in a meandering way. If it was cleaned up, and focused on the perils of ecological research and framing the spectrum of human/health/population research it would improve the paper.

Response: The following text has been added:
Ecological research can address important issues that cannot be easily addressed by other study designs. They are frequently used where alternative study designs are not possible (eg, randomized control trials), such as when investigating the effect of geographical factors on disease incidence. Our approach to BT consumption presents a number of limitations like all ecological studies because factors other than dietary habits may be the most important determinants of variations in diabetes prevalence across communities. For example, it is possible that other unmeasured confounding factors (eg, genetic differences)may explain some of the observed regional variations. Due to the large number of potential determinants of diabetes prevalence, including patient-, physician-, hospital-, and community-related variables, it is difficult to identify with certainty all of the causes of the regional variations of diabetes prevalence, and additional follow-up studies should be considered to confirm the hypotheses generated by this type of study.
A number of biological, physiological and epidemiological studies have provided evidence linking BT consumption and glucose metabolism 25, 29, 30, 36-38, 45, 46. However, a large-scale, longitudinal,

prospective case-control study comparing high BT consumption versus no consumption and diabetes prevalence would be useful to confirm these findings.

Reviewer 2: Lesley Butler
Colorado State University, Department of Environmental & radiological,
Health Sciences

Comment 26: Inappropriate application of multiple correspondence analysis (MCA). This reviewer is not familiar with MCA, so I looked up the reference provided on p.5, line 53 (#13 Briand, 2009). Unfortunately, this reference provides an example of an application of MCA. A reference for the method itself is needed. Using reference #13 to judge the application of MCA in the present analysis led this reviewer to question it's appropriateness. For example in the Briand, et al study, "The MCA modeling method was specifically adapted to reduce the five exposure variables to one aggregated exposure indicator…MCA provides a standardized way to reduce complexity." This is quite different from the use of MCA used in the present analysis, where MCA was used to identify a correlation between one exposure variable and 5 outcome variables. From the documentation provided, MCA does not appear to be a method that is used to pull out an individual exposure-disease relationship, as it was used in the present analyses.

Response: MCA calculates the distances between the points in m-dimensional space to summarize all information concerning the similarities between the m variables. A lower-dimensional space can then be defined, in which to position the row points in such a way that it retains most of the original information. The Briand et al study used MCA to project the original information onto a single axis. This study used MCA to project the original information onto 2 and 3 axes. Vector positioning then demonstrated potential correlations between variables. These 2 studies therefore used MCA appropriately, according to their initial objectives.

Commment 27: Unspecific citations for the two data sources (p.4-5). The market research resource for the black tea intake is not publically available (reference #7). I could not locate information about this data source from the reference information provided. As for the WHO disease prevalence information, this reference was also inadequate (#8). There are several WHO surveys listed on the website provided. Which data was used in the present analysis? It appears that the WHO has data on many regions worldwide, but that this information is not country specific, unlike the black tea information.

Response: The citations of the data sources are correct but the datasets are not publicly available. The Euromonitor dataset is available to subscribers under strict terms and conditions. Concerning epidemiological data from the World Health Survey, the web reference is adequate and most data are available online in the form of specific country reports, which can be downloaded. However, it is true that only researchers collaborating on a regular basis with WHO have access to the full dataset.

Comment 28: In the Discussion, the authors go back and forth between providing evidence for green and black tea intake/compounds and disease associations. It appears that the authors incorrectly consider evidence for green tea to be synonymous with evidence for black tea. For example (p.8, lines 43-45), the health benefits of green tea are used to support the author's findings for black tea. The references provided on p. 8, line 54 mix references with findings for green and black tea. On page 9, lines 36-40 describe studies with findings (reference 44 and 45) supporting the beneficial

health effects of green tea (or green tea catechins), but the authors simply state “tea” without specifying type. On line 33 (p. 9), black tea should be specified for the findings reported for reference #40.

Response: As explained, black tea is the result of a transformation process of green tea based on fermentation. Both types of tea contain flavonoids in different compositions. Tea studies are sometimes published without any details concerning the type of tea (Feng 2010; de Mejia 2009, Walsh 1997). We therefore decided to provide references to studies conducted on green tea and tea in general.

Comment 29: In the Background, a rationale is needed for the selection of the 5 disease outcomes chosen. Only a rationale for diabetes is provided, although theauthors state that the relationship between black tea and diabetes was not an a priori hypothesis (p.8, second paragraph).

Response: The sentence has been revised as follows:
Various studies, especially in Asian populations, confirm that flavonoids present in green tea could reduce fat absorption in the gut, may promote fat oxidation in tissues and may increase energy expenditure 47.

Comment 30: In the Methods, references are needed for the “correlation circle” paragraph/method (p.6).

Response: A reference has been added:
Everitt B, Dunn G, Applied Multivariate data analysis, Lavoisier, 2d ed 2001, 320p

Comment 31: In the Results, provide a graph that plots each country/region by amount of black tea intake (x-axis) by the prevalence of diabetes (y-axis). Do this for each outcome. This more common graphical presentation would help the audience grasp what is being shown in the less common correlation circle.

Response: The suggested graphs have been constructed and one is presented below. However, this graph does not appear to be sufficiently clear to be published due to overlapping of the large number of countries labels. Furthermore, new figures 1 and 2 provide the same information in the form of bar charts, which are easier to understand. We will therefore let the publisher decide whether or not to publish such graph.

Comment 32: The heterogeneity of all of the diseases (especially cancer and infectious diseases) is a limitation and may be responsible for the lack of a statistical relationship with black tea. This issue needs to be mentioned in the Discussion.

Response: The following sentence has been added:
Some of the selected health indicators represent a group of diseases, such as infectious diseases (tuberculosis and HIV) and cancer. The heterogeneity of these indicators can make it difficult to establish any potential statistical relationships. Although more homogeneous, health indicators such as diabetes depend on diagnostic criteria, which can vary from country to country.

VERSION 2 – REVIEW

REVIEWER	Huxley, Rachel The George Institute, Nutrition and Lifestyle Division
REVIEW RETURNED	07-Mar-2012

GENERAL COMMENTS	This is essentially an ecological study designed to look at the correlation between consumption of black tea and key health indicators in 50 countries. The main finding of this paper is a significant negative linear relationship between black tea consumption with diabetes prevalence (but not with CVD, respiratory disease, infectious disease or cancer) suggesting the a specific effect. This reviewer does have some significant concerns regarding the manuscript. Ecological studies are most useful for establishing some evidence for a novel causative risk factor in disease (Ancel Keys seven country study on fat intake and CHD being one of the most famous). However, because these studies are done at the population level they are often subject to confounding at the individual level (in this case the relationship between tea consumption and diabetes could be confounded by body mass index) they are at the lowest level of the evidence hierarchy in epidemiology with the next stage being case-control studies, cohort studies and finally randomised trials. The hypothesis that tea consumption might be protective against diabetes has been well documented in the literature (e.g. Huxley et al. Archives of Internal Medicine 2009) and therefore it is questionable what this ecological analysis can add to the evidence-base, indeed it seems to be a retrograde step to utilise this study design. Rather, what is needed now is not a "large-scale longitudinal prospective case-control study" (which is an oxymoron - one can't have a longitudinal case-control study) as the authors state on page 11, but a large-scale randomised controlled trial of tea consumption and diabetes risk. Indeed there have been numerous small trials on this topic that disappointingly, have all produced null findings.

REVIEWER	Illhoi Yoo Assistant Professor of Health Informatics Department of Health Management and Informatics, University of Missouri School of Medicine
REVIEW RETURNED	19-Jun-2012

THE STUDY	No participant is involved in the study. No patient or patient related data is involved in this study. This study does not require following a standard of reporting such as CONSORT.
RESULTS & CONCLUSIONS	The study demonstrates that diabetes prevalence decreases with the increase of black tea consumption. However, it does not really show that chances of diabetes are low within the group of people drinking black tea. See more comments in the section “comments for authors”
REPORTING & ETHICS	This study does not require following a standard of reporting such as CONSORT.
GENERAL COMMENTS	1. The title of the article starts with “data mining approach”. It seems that because the authors used PCA, they claim they used a data

	mining approach. However, PCA is not regarded as data mining but statistics. Thus, the title should be changed (Minor essential revision) 2. Though the objective of the study is to investigate potential statistical relationships with BT consumption and five health indicators (i.e. respiratory diseases, infectious diseases, cancer, cardiovascular diseases and diabetes), only the impact of diabetes is discussed in the "Background" section. I suggest that the background section should be expanded focusing on all of the five health indicators. (Minor essential revision) 3. A sample of the data should be included in the manuscript so that the study can be better understood. (Minor essential revision) 4. As the number of populations of different countries is not same, the tea consumption and epidemiological data should be normalized based on a fixed number of population. (Major compulsory revision) 5. Page 3, line 46, "A frequent criticism of the use f data mining …." should be "A frequent criticism of the use of data mining …." (Minor essential revision) 6. Page 5, line 24-34, the discussion of data mining is misleading. Please remove this paragraph from the manuscript. (Minor essential revision) 7. Page 5, line 33-34, in place of references 11-14, the following two review articles on data mining in healthcare would greatly benefit the statement. (Discretionary revision) + Yoo I, Alafaireet P, Marinov M, Pena-Hernandez K, Gopidi R, Chang J-F, Hua L: Data Mining in Healthcare and Biomedicine: A Survey of the Literature. Journal of medical systems 2011. + Marinov M, Mosa ASM, Yoo I, Boren SA: Data-mining technologies for diabetes: a systematic review. Journal of diabetes science and technology 2011, 5:1549-56. 8. Page 5, line 39-40, please define "synthetic dimension" and specify all three dimensions. (Minor essential revision) 9. If possible, please provide a 3D representation of figure 3. (Minor essential revision) 10. Page 7, line 55, "what does the number 42 represent?" (Minor Essential revision) 11. The linear regression equation in page 8 (line 4-6) has a large intercept value. Discuss the results in terms of this large intercept. (Minor revision) 12. The limitations of the study are discussed at the beginning of "discussion" section. Please move the limitations under the heading "Limitations". (Minor revision) 13. The discussion section should discuss the regression line (figure 4) and why density of points in figure 4 is higher in certain regions whereas density decreases with the increase of tea consumption amount. Data normalization may change the result significantly. The similar analysis after data normalization is recommended. (Major compulsory revision) 14. Scatter plots of BT consumption vs. five health indicators will be useful to see if there is any linear or curvilinear relation exists between BT consumption and other four health indicators. (Minor essential revision) 15. Page 9, line 18-23, nine references (17-25) have been used to support the statement. Instead, it is recommended to find systematic review articles (if any). (Minor revision) 16. Figure 2 is condensed and hard to follow. (Minor essential revisions)

REVIEWER	Heiner Boeing Nuthetal Germany
REVIEW RETURNED	19-Jul-2012

GENERAL COMMENTS	This is a manuscript that use the new term of data mining for the old approach of ecological correlations. The hypothesis is that black tee consumption and diabetes prevalence correlate with each other. In this way it is not an exploratory data mining approach but a hypothesis driven analysis on which criteria needs to be applied that reflect the current standard of epidemiological research on this question. In the past this type of study design has been used to generate first hypothesis on the role of diet (about 1970). It had been criticized that the bias of ecological fallacy could not been excluded when using aggregate data compared to data in which the exposure status of the subjects which get the disease is known. This does not mean that this study design is not useful anymore but that the design could only been used in a very early phase of hypothesis generation. Further, in case of this manuscript, I am in doubt that we have reliable data on diabetes incidence and prevalence worldwide, that prevalence data instead of incidence data are a useful parameter, that the small slope of the correlation is meaningful, and that there is no conflict of interest. The conclusion of this paper is far beyond the type of evidence being generated.

VERSION 2 – AUTHOR RESPONSE

Comments from Reviewer 1
Rachel Huxley, Associate Professor, University of Minnesota

Comment 2:

This is essentially an ecological study designed to look at the correlation between consumption of black tea and key health indicators in 50 countries. The main finding of this paper is a significant negative linear relationship between black tea consumption with diabetes prevalence (but not with CVD, respiratory disease, infectious disease or cancer) suggesting the a specific effect. This reviewer does have some significant concerns regarding the manuscript. Ecological studies are most useful for establishing some evidence for a novel causative risk factor in disease (Ancel Keys seven country study on fat intake and CHD being one of the most famous). However, because these studies are done at the population level they are often subject to confounding at the individual level (in this case the relationship between tea consumption and diabetes could be confounded by body mass index) they are at the lowest level of the evidence hierarchy in epidemiology with the next stage being case-control studies, cohort studies and finally randomised trials. The hypothesis that tea consumption might be protective against diabetes has been well documented in the literature (e.g. Huxley et al. Archives of Internal Medicine 2009) and therefore it is questionable what this ecological analysis can add to the evidence-base, indeed it seems to be a retrograde step to utilise this study design. Rather, what is needed now is not a "large-scale longitudinal prospective case-control study" (which is an oxymoron - one can't have a longitudinal case-control study) as the authors state on page 11, but a large-scale randomised controlled trial of tea consumption and diabetes risk. Indeed there have been numerous small trials on this topic that disappointingly, have all produced null findings.

Answer:
Although numerous potential health benefits of Black tea have already been published, the value of this study is to establish which of the 5 key health indicators would be potentially correlated with Black Tea consumption using both ecological and data mining approaches. Correlation established by normative PCA is a new approach, confirming the strong correlation. This correlation has never been previously established worldwide by any existing biological or field studies and can be considered to be an additional piece of evidence based on a new perspective. Of particular interest is the recent publication of a similar ecological study in Europe (Tea Consumption and Incidence of Type 2 Diabetes in Europe: The EPIC-InterAct Case-Cohort Study. *PLoS One.* 2012(7):5) confirming our results and establishing that this approach provides relevant new information in addition to existing biological and field studies and reviews such as the meta-analysis published by Huxley et al (2009). Global epidemiological and tea consumption data are very complex to collect and our worldwide perspective confirms and complements most of the available local studies

The following text has been added to the discussion section:

The main criticism of this approach is the ecological fallacy, corresponding to a logical fallacy in interpretation of the observed correlations at the population level, assuming that they can be applied at the individual level. It is well known that statistics that accurately describe group characteristics do not necessarily apply to individuals within that group. Our study on black tea does not comprise any potential logical fallacy, as it was not used as a basis for any individual assumptions. However, when strong associations are observed , the results of ecological studies have provided numerous assumptions that have been subsequently confirmed by experimental studies. One of the best known ecological studies was that published by Keys in 1980 concerning the relationship with dietary habits and coronary heart disease in 7 selected countries. The results of what later became known as the “Seven Countries Study” appeared to show that serum cholesterol was strongly related to coronary heart disease mortality at both the population and individual levels, leading to US government dietetic guidelines. Other ecological studies have significantly contributed to scientific knowledge and public health interventions, such as the relationship between lung cancer and tobacco, which has been confirmed by numerous studies (Didkowska, 2005).

The following references have been added:
Didkowska J, Manczuk M, McNeill A, Powles J, Zatonski W. Lung cancer mortality at ages 35-54 in the European Union: ecological study of evolving tobacco epidemics. BMJ. 2005 Jul 23;331(7510):189-91.

Keys A. Seven Countries: A Multivariate Analysis of Death and Coronary Heart Disease. 1980, Harvard University Press

The following sentence:

However, a large-scale, longitudinal, prospective case-control study comparing high BT consumption versus no consumption and diabetes prevalence would be useful to confirm these findings.

has been replaced by:

"However, a large-scale randomised controlled trial of tea consumption and diabetes risk would be useful to confirm these findings."

Comments from Reviewer 2
Illhoi Yoo, Assistant Professor of Health Informatics
Department of Health Management and Informatics, University of Missouri School of Medicine

Comment 3: The title of the article starts with "data mining approach". It seems that because the authors used PCA, they claim they used a data mining approach. However, PCA is not regarded as data mining but statistics. Thus, the title should be changed (Minor essential revision)

Response: Data mining used methods overlapping the fields of statistics, artificial intelligence and machine learning in order to extract meaningful information from a dataset. Because PCA techniques have been developed in Europe in order to manage large databases, PCA is widely used in Europe in the context of data mining. It is true that many information technology specialists use machine learning and artificial intelligence techniques more than statistics for data mining purposes. In any case, we have proposed a new title, which would be more suitable for both sides of the Atlantic Ocean.

The title has been changed as follows:

Relationships Between Black Tea Consumption and Key Health Indicators in the World: an Ecological Study

Comment 4: Though the objective of the study is to investigate potential statistical relationships with BT consumption and five health indicators (i.e. respiratory, diseases, infectious diseases, cancer, cardiovascular diseases and diabetes), only the impact of diabetes is discussed in the "Background" section. I suggest that the background section should be expanded focusing on all of the five health indicators.

Response:

The following text has been added to the background section:

"Most recent studies use multidisciplinary approaches including epidemiology, field studies, and laboratory research in animal models, mostly for respiratory diseases, infectious diseases, heart diseases, various types of cancers and diabetes, as well as *in vitro* experiments. In respiratory diseases, several tea components have been established to be effective in airway diseases. Tea catechin polyphenols seems to be effective to improve inflammation of obliterative airway disease (Liang, 2011), protect against oxidative damage and apoptosis in human bronchial epithelial cells induced by tobacco (Qing, 2010) or attenuate oxidative responses to intermittent hypoxia (Burckardt, 2008). In infectious diseases, herbal products have gained considerable interest among pharmaceutical companies and consumers due to the minimal perceived side effects associated with these products. Several antimicrobial activities have been attributed to tea flavonoids. Catechins appear to have virucidal and virustatic actions (Marathe, 2012) and exert a protective activity against *Vibrio cholerae* (Toda, 1992). However, research into the potential beneficial effects of tea appears to be most active in the field of cardiovascular disease, in view of the number of recent publications in this field. Most of these publications tend to confirm that tea catechins would exert cardioprotective effects via various mechanisms including reversal of endothelial dysfunctions, reduction of inflammatory biomarkers, and antioxidant, antiplatelet and antiproliferative effects (Islam, 2012). Moreover, dietary consumption of tea catechins would have beneficial effects on blood pressure and lipid parameters (Son, 2012; Hodgson, 2012). Similarly, a number of studies have focused on the potential effects of tea in cancer. Biochemical and biological studies, prospective cohort studies and double-blind randomized clinical prevention trials tend to show convergent results for the beneficial

preventive effects of tea components in various cancers such as hepatocellular carcinoma, skin, prostate, lung or colorectal cancer (Fujiki, 2012)."

The following references have been added:

Liang OD, Kleibrink BE, Schuette-Nuetgen K, Khatwa UU, Mfarrej B, Subramaniam M. Green tea epigallo-catechin-galleate ameliorates the development of obliterative airway disease. Exp Lung Res. 2011, 37(7):435-44

Qing C, Chen P, Xiang X. Effect of tea polyphenols on oxidative damage and apoptosis in human bronchial epithelial cells induced by low-dose cigarette smoke condensate. Zhong Nan Da Xue Xue Bao Yi Xue Ban. 2010;35(2):123-8

Burckhardt IC, Gozal D, Dayyat E, Cheng Y, Li RC, Goldbart AD, Row BW. Green tea catechin polyphenols attenuate behavioral and oxidative responses to intermittent hypoxia. Am J Respir Crit Care Med. 2008, 177(10)1135-41

Marathe SA, Datey AA, Chakravortty D. Herbal Cocktail as Anti-infective: Promising Therapeutic for the Treatment of Viral Diseases. Recent Pat Antiinfect Drug Discov. 2012;7(2):123-32

Toda M, Okubo S, Ikigai H, Suzuki T, Suzuki Y, Hara Y, Shimamura T. The protective activity of tea catechins against experimental infection by Vibrio cholerae O1, Microbiol Immunol. 1992;36(9):999-1001

Islam MA, Cardiovascular effects of green tea catechins: progress and promise. Recent Pat Cardiovasc Drug Discov. 2012;7(2):88-99

Son JT, Lee E, Effects of green tea ingestion on postprandial drops in blood pressure in older adults. J Gerontol Nurs. 2012; 38(3):30-8

Hodgson JM, Puddey IB, Woodman RJ, Mulder TP, Fuchs D, Scott K, Croft KD, Effects of black tea on blood pressure: a randomized controlled trial. Arch Intern Med. 2012;172(2):186-8

Fujiki H, Imai K, Nakachi K, Shimizu M, Moriwaki H, Suganuma M, Challenging the effectiveness of green tea in primary and tertiary cancer prevention. J Cancer Res Clin Oncol. 2012;138(8):1259-70

Comment 5: A sample of the data should be included in the manuscript so that the study can be better understood.

Response:
The following table has been added:

Table 1: Sample of the dataset presenting the five key health indicators (rate per 100,000 inhabitants) and tea consumption in 8 countries (kg per 100,000 inhabitants).

Country	Respiratory diseases	Infectious diseases (TB, HIV)	Cancers	Cardiovascular diseases	Diabetes	Black Tea consumption
Indonesia	2063	306	776	1063	5639	30710
Romania	2237	228	2361	3399	6772	590
Russia	2394	748	2078	4113	4050	106680
Hungary	2505	62	2204	4685	5927	11270
Ukraine	2552	857	2245	4630	4612	32290
Turkey	2931	48	1271	1579	3326	166310
Egypt	3121	40	615	1316	3979	95910
Saudi Arabia	3221	54	353	914	4257	57020

Comment 6: As the number of populations of different countries is not same, the tea consumption and epidemiological data should be normalized based on a fixed number of population.

Response: Prevalences and black tea consumption have both been expressed per 100,000 inhabitants, to normalize the results to the size of the population.

Comments 7: Page 3, line 46, "A frequent criticism of the use f data mining …." should be "A frequent criticism of the use of data mining …."

Response: the typo has been corrected.

Comments 8: Page 5, line 24-34, the discussion of data mining is misleading. Please remove this paragraph from the manuscript.

Response: This paragraph has been deleted from the manuscript.

Comments 9: Page 5, line 33-34, in place of references 11-14, the following two review articles on data mining in healthcare would greatly benefit the statement.
Yoo I, Alafaireet P, Marinov M, Pena-Hernandez K, Gopidi R, Chang J-F, Hua L: Data Mining in Healthcare and Biomedicine: A Survey of the Literature. Journal of medical systems 2011.
Marinov M, Mosa ASM, Yoo I, Boren SA: Data-mining technologies for diabetes: a systematic review. Journal of diabetes science and technology 2011, 5:1549-56.

Response: The two articles suggested by the reviewer have been added to the references.

Comment 10: Page 5, line 39-40, please define “synthetic dimension” and specify all three dimensions

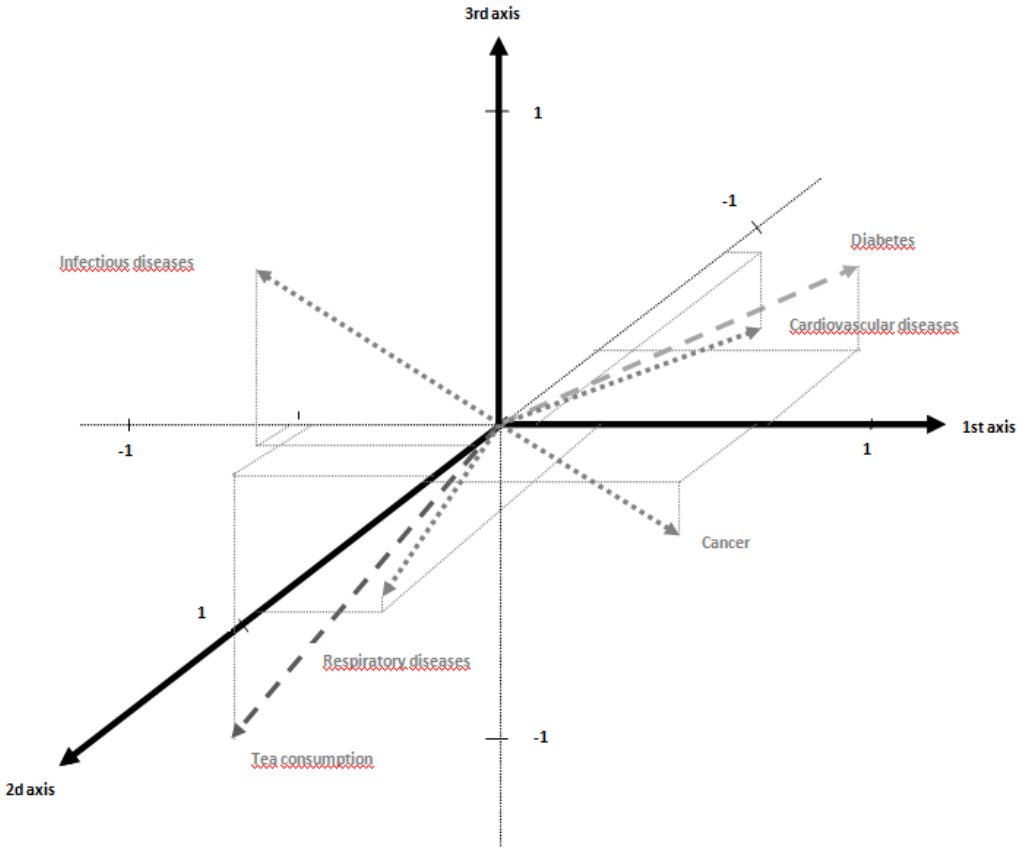
Response:
Each factorial axis (dimension) in the space defined by the subjects (in this case, the countries) is defined by a main vector representing a statistical variable, which is not measured but calculated (latent statistical variable) based on the fact that the residual variance of the explanation of the "measured variables" by the latent statistical variable is minimum, or (which is equivalent) that the explicative variance is maximum. This represents the best possible mathematical projection of the 6 dimensions represented by the 6 variables. Factorial axes are numbered by decreasing order of variance. The first axis is that which best explains the measured variables, followed by the second axis, etc. Each factorial axis can be considered to represent a cluster of the calculated variables.

The following text:
"firstly, a "calibrated principal component analysis" (PCA) was used to segment the database composed of 6 variables (BT consumption and the 5 health indicators) into 3 synthetic dimensions; secondly, the 6 variables were represented as vectors in a "correlation circle" to study potential positive or negative correlations; finally, a linear correlation model was tested on selected variables."

has been changed as follows:
"firstly, a "calibrated principal component analysis" (PCA) was used to segment the database composed of 6 variables (BT consumption and the 5 health indicators) into 3 synthetic dimensions represented by 3 axes which can be considered to be the mathematical projection of the 6 dimensions defined by the 6 variables in 3 dimensions; secondly, the 6 variables were represented as vectors in a "correlation circle" to study potential positive or negative correlations; finally, a linear correlation model was tested on selected variables."

Comment 11: If possible, please provide a 3D representation of figure 3.

Response: Figure 3 has been redesigned in 3D and is now entitled
"Figure 3: Three dimensional correlation circle of 5 health indicators and BT consumption*"



Comments 12: Page 7, line 55, "what does the number 42 represent?"

Response: 42 is the number of countries included in the linear regression. As this information is already mentioned in the last paragraph and as this figure in brackets appears to be unclear, it has been deleted from the end of this sentence.

Comments 13: The linear regression equation in page 8 (line 4-6) has a large intercept value. Discuss the results in terms of this large intercept.

Response:
The following text has been added to the results section:
The y-coordinate of the point at which the regression line intersects the y-axis (intercept) can be considered to correspond to the average prevalence of diabetes in a country in which BT consumption is unknown (6,173 cases per 100,000 inhabitants).

Comments 14: The limitations of the study are discussed at the beginning of "discussion" section. Please move the limitations under the heading "Limitations"

Response: Subheadings have been included in the discussion section starting with a "Limitations" sub-section.

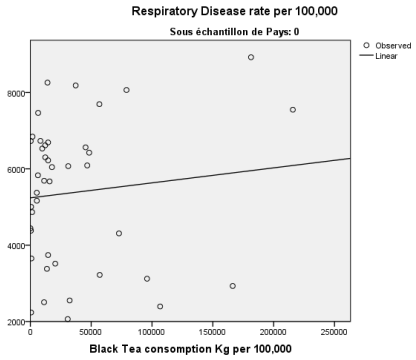
Comments 15: The discussion section should discuss the regression line (figure 4) and why density of points in figure 4 is higher in certain regions whereas density decreases with the

increase of tea consumption amount. Data normalization may change the result significantly. The similar analysis after data normalization is recommended

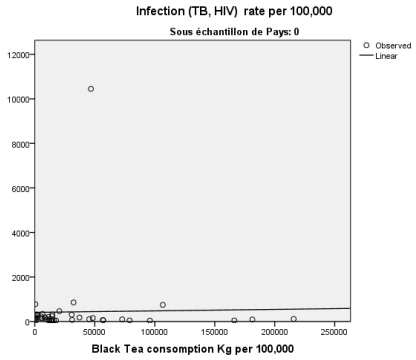
Response:
See response to comment 6. All variables have been normalized per 100,000 inhabitants.

Comments 16: Scatter plots of BT consumption vs. five health indicators will be useful to see if there is any linear or curvilinear relation exists between BT consumption and other four health indicators.

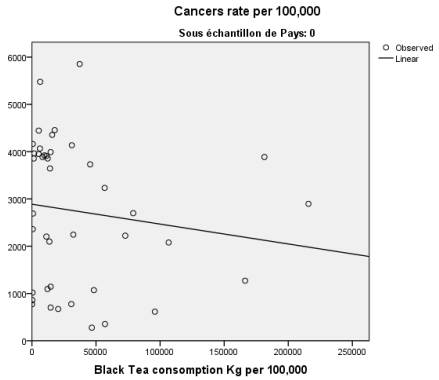
Response:
Scatter plots of BT consumption versus each of the five health indicators are presented below. The only significant linear relationship was the relationship between BT consumption and Diabetes prevalence.



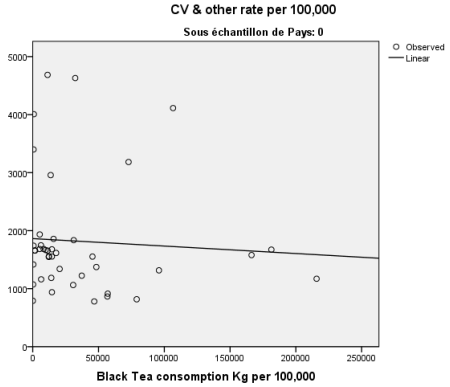
$r^2 = 0.011$
$F = 0.450$
$p = 0.506$



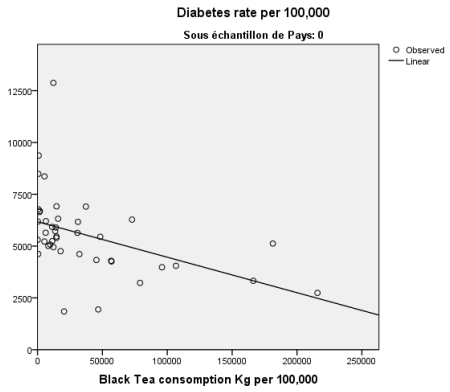
$r^2 = 0.000$
$F = 0.018$
$p = 0.895$



$r^2 = 0.019$
$F = 0.776$
$p = 0.383$



$r^2 = 0.004$
$F = 0.166$
$p = 0.686$



r^2 = 0.199
F = 9.947
p = 0.003

Comments 17: Page 9, line 18-23, nine references (17-25) have been used to support the statement. Instead, it is recommended to find systematic review articles (if any).

Answer: No systematic reviews were found in Medline/Pubmed databases.

Comments 18: Figure 2 is condensed and hard to follow

Answer: Figure 2 has been replaced by a sample of the dataset (Table 1).

<u>Comments from Reviewer 3 : Heiner Boeing - German Institute of Human Nutrition (DIFE) - Department of Epidemiology</u>

Comment 19: This is a manuscript that use the new term of data mining for the old approach of ecological correlations. The hypothesis is that black tee consumption and diabetes prevalence correlate with each other. In this way it is not an exploratory data mining approach but a hypothesis driven analysis on which criteria needs to be applied that reflect the current standard of epidemiological research on this question.
In the past this type of study design has been used to generate first hypothesis on the role of diet (about 1970). It had been criticized that the bias of ecological fallacy could not been excluded when using aggregate data compared to data in which the exposure status of the subjects which get the disease is known. This does not mean that this study design is not useful anymore but that the design could only been used in a very early phase of hypothesis generation. Further, in case of this manuscript, I am in doubt that we have reliable data on diabetes incidence and prevalence worldwide, that prevalence data instead of incidence data are a useful parameter, that the small slope of the correlation is meaningful, and that there is no conflict of interest. The conclusion of this paper is far beyond the type of evidence being generated

Response: See response to comment 1.
According to the old approach of ecological correlations, correlation tests should be carried out directly on the variables of the dataset. Our data mining approach can be used to study the structure of the dataset before performing any correlation tests, which is an innovative approach. A new ecological study published in PLoS One in May 2012 "Tea Consumption and Incidence of Type 2 Diabetes in Europe: The EPIC-InterAct Case-Cohort Study" by the InterAct Consortium not only confirms our findings (a linear inverse association was observed between tea consumption and

incidence of type 2 diabetes in Europe) but also established that ecological studies can provide original information, even beyond the early hypothesis phase.
Epidemiological data are extracted from the World Health Survey (WHS), which collects epidemiological data worldwide under the supervision of the World Health Organization. These data are used by a number of epidemiological analytical studies.
The only declared conflict of interest concerns Prof. Berger who is an employee at Unilever and who provided the tea consumption world dataset for this research. Prevalence data versus incidence are useful when investigating chronic diseases. The new reference mentioned above (InterAct consortium, 2012) used incidence data and confirmed our findings with prevalence data. The slope coefficient is highly significantly different from 0 and contributes to confirm the robustness of the linear relationship, regardless of the slope. The authors confirm that this study is totally independent, has not been funded by any sponsors and was not subject to any potential restrictions or clearances, regardless of the results.