Negative Correlation Between Black Tea Consumption and Diabetes Prevalence in the World

Journal:	*BMJ Open*
Manuscript ID:	bmjopen-2011-000648
Article Type:	Research
Date Submitted by the Author:	19-Nov-2011
Complete List of Authors:	BERESNIAK, ARIEL; Data Mining International Duru, Gerard; Data Mining International, Research Berger, Genevieve; Unilever, R&D Bremond-Gignac, Dominique; Amiens University Hospital, INSERM UMRS 968
Primary Subject Heading:	Epidemiology
Secondary Subject Heading:	Diabetes and endocrinology, Nutrition and metabolism, Public health
Keywords:	DIABETES & ENDOCRINOLOGY, NUTRITION & DIETETICS, EPIDEMIOLOGY

Negative Correlation Between Black Tea Consumption and Diabetes Prevalence in the World

Correlation Between Black Tea Consumption and Diabetes Prevalence

Ariel Beresniak, MD, MPH, PhD, Data Mining International, Geneva, Switzerland
Gerard Duru, PhD, Data Mining International; Geneva, Switzerland
Genevieve Berger, MD, PhD, Unilever, London, UK
Dominique Bremond-Gignac, MD, PhD, Amiens University Hospital, Amiens, INSERM UMRS 968, Paris VI University, France

Ariel Beresniak, MD, MPH, PhD
Data Mining International
Route de l'Aeroport, 29-31
CP221
CH-1215 Geneva 15
Switzerland
Phone: + 41 22 799 34 00 Fax . + 41 22 788 38 50
aberesniak@datamining-international.com

Funding statements : This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors".

Competing interest statement : 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 excepting G.B.who is Chief R&D Officer at Unilever and provided access to global tea consumption data, no other relationships or activities that could appear to have influenced the submitted work.

Acknowledgements: The authors thank Professor Paul Zimmet, Director Emeritus, Baker IDI Heart and Diabetes Institute, for reviewing the manuscript and making comment".

Data sharing statements: This study used epidemiological data from WHO World Health Surveys.

Individual contributions: The work presented here was carried out in collaboration between all authors. A.B. conceived the study aims and design. G.B. contributed to the data collection. A.B. and G.D performed the analysis. A.B, G.D and D.B. interpreted the results. A.B drafted the manuscript. All authors have contributed to, seen and approved the manuscript.

Abstract

Objective: The objective of this study was to investigate a possible correlation between black tea consumption and key health indicators in the world, including diabetes.

Methodology: A systematic data mining approach was carried out on black tea consumption data and five key health epidemiological indicators from the World Health Survey (WHO): respiratory diseases, infectious diseases, cancer, cardiovascular diseases and diabetes. The methodological approach included 3 phases: firstly, a "calibrated principal component analysis" was used to segment the database composed of 6 variables (black tea consumption and 5 health indicators) into 3 dimensions; secondly, the 6 variables were represented as vectors in a projected "correlation circle" to study potential positive or negative correlations; lastly, a linear correlation model was tested on selected variables.

Results: Principal component analysis established a very high contribution of the black tea consumption parameter on the 3rd axis (81%). The correlation circle represented the "black tea" vector strictly opposite to the "diabetes prevalence" vector, suggesting a negative statistical correlation. A linear correlation model then confirmed a significant statistical correlation between high black tea consumption and low diabetes prevalence.

Conclusion: This innovative study establishes, for the first time, a linear statistical correlation between high black tea consumption and low diabetes prevalence in the world. Although the objective of this analysis was not to demonstrate any direct cause-and-effect relationship, these results are consistent with biological and physiological studies conducted on the potential effect of black tea on diabetes and obesity. Further epidemiological research is necessary to investigate the causality.

Article summary

Article focus :

This study investigates potential statistical relationships between Black Tea consumption and a selection of key health indicators in 50 countries

Key messages :

- A linear statistical correlation between high black tea consumption and low diabetes prevalence has been established.
- These results are consistent with biological and physiological studies conducted on the potential effect of black tea on diabetes and obesity.
- These results should support further causality research regarding the health benefits of BT consumption on type 2 diabetes prevalence in the world.

Strengths and limitations :

- These original study results are consistent with biological and physiological studies conducted on the potential effect of black tea on diabetes and obesity. We believe that this multidimensional approach provides valuable additional scientific information, which is why our findings establishing a strong correlation between high BT consumption and low diabetes prevalence should be considered as a contribution to existing biological, physiological and epidemiological studies conducted on tea consumption, diabetes and obesity.
- Diabetes prevalence data were obtained from the World Health Survey implemented by the World Health organization, which constitutes an official source of key morbidity indicators around the world. However, the quality of data collection can be expected to be heterogeneous around the world and diabetes diagnostic criteria can vary from country to country.
- Another important concern is the interpretation of the established statistical relationship between BT consumption and diabetes prevalence. The number of factors contributing to the growth of diabetes and obesity in the world confirm that "correlation does not imply causality ", and that a significant linear correlation between BT consumption and diabetes prevalence does not imply that low BT consumption could *cause* diabetes. A correlation can only indicate a potential direct or indirect possible cause, which then needs to be further investigated.
- A frequent criticism of using data-mining was based on the confusion between *data-mining* and *data-dredging* techniques. While a data-mining approach is based on searching for combinations of variables that might show potential correlations, data-dredging can generate misleading results. When a number of hypotheses are tested, it is expected that some will falsely appear to be statistically significant, since every database can contain potential random correlations. A robust data mining approach must therefore always be based on a clear research strategy and a limited number of relevant meaningful assumptions.

Background

An almost 6-fold increase in the number of people with diabetes has been observed over the last few decades. The International Diabetes Federation (IDF) reports that the number of people with diabetes will escalate from 285 million to 438 million between 2010 and 2030 [1] and the number of persons with IGT will increase from 344 to 472 million. By 2030, there will be over 900 million people worldwide with diabetes or at high risk of diabetes. Diabetes confers about a two-fold excess risk for a wide range of vascular diseases [2]. Furthermore, diabetic retinopathy is a common and specific microvascular complication of diabetes, and remains the leading cause of preventable blindness in working-aged people [3]. With one of the highest prevalences of all human diseases, diabetes is now a global epidemic with devastating health, social and economic consequences [4]. In certain ethnic groups, such as Asian populations, diabetes develops at a younger age than in Caucasian populations. Several distinctive features are apparent in the pathogenic factors for diabetes and their thresholds in Asian populations [5]. In conjunction with genetic susceptibility, type 2 diabetes is brought on by environmental and behavioural factors such as a sedentary lifestyle, overly rich nutrition and obesity and results in a huge economic burden [6]. It could therefore be interesting to investigate some key dietary habits in relation with life style and health effects at a global level. For example, the positive health effects of black tea (BT) have been observed for centuries.

Considering the variability of dietary habits and BT consumption according to longitude, no epidemiological study has yet investigated any potential statistical relationship between key health indicators and worldwide BT consumption. Potential correlations between BT consumption and epidemiological data around the world could therefore be investigate by deploying a data mining approach and advanced exploratory statistical methods.

The objective of this original research was to investigate potential statistical relationships between BT consumption and a selection of key health indicators in 50 countries.

Material and Method

Data sources

BT consumption data were derived from a specific international trade survey compiling sales data conducted by Euromonitor, an independent agency specialized in market research [7]. Consumption data expressed in kilograms per capita were available for the following 50 countries: Argentina, Australia, Austria, Belgium, Brazil, Bulgaria, Canada, Chile, China, Colombia, Czech Republic, Denmark, Egypt, Finland, France, Germany, Greece, Hungary, India, Indonesia, Ireland, Israel, Italy, Japan, Malaysia, Mexico, Morocco, Netherlands, New Zealand, Norway, Philippines, Poland, Portugal, Romania, Russia, Saudi Arabia, Singapore, Slovakia, South Africa, South Korea, Spain, Sweden, Switzerland, Thailand, Turkey, Ukraine, United Kingdom, USA, Venezuela, Vietnam.

Epidemiological data were derived from a specific analysis of the World Health Survey (WHS) conducted by the World Health Organization (WHO). Each year, the WHS compiles comprehensive baseline information on the health of populations and health system outcomes [8]. Five key health indicators were selected in 50 countries in both men and women for all age groups: prevalence of respiratory diseases, prevalence of infectious diseases (tuberculosis and HIV), prevalence of cancer, prevalence of cardiovascular diseases and prevalence of diabetes.

Methods

Data analyses were based on a systematic data-mining approach. Data-mining (sometimes called data or knowledge discovery) is generally defined as the process of analysing data from different perspectives and summarising these data into meaningful information. This approach is useful to analyse data derived from different dimensions or perspectives and to detect potential relationships between variables. Technically, data-mining consists of discovering specific correlations or patterns in large relational databases. Data-mining combines methods from statistics and artificial intelligence with database management and is considered to be an increasingly important tool. It is currently used in a wide range of scientific applications in health [9-12].

In this study, the data-mining approach used 3 phases: firstly, a "calibrated principal component analyses" (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.

Normative principal component analysis (PCA)

PCA is a mathematical procedure that uses mathematical projections to convert a set of *n* possibly correlated variables representing *n* dimensions into a smaller number of dimensions called "principal components" classically represented in 2 or 3 axes F1, F2, F3. The projections use orthogonal transformations defined in such a way that the first principal component (first axis) has the highest possible variance in order to synthesize most of the initial information. The main objective of PCA is to reduce the dimensionality of the data set. PCA is often presented as a technique of factor analysis for quantitative variables. Multiple Correspondence Analysis (MCA) is another type of factor analysis for quantitative, qualitative and categorical variables and is useful to conduct multi-criteria analyses such as multi-criteria risk assessment [13]. A "normative PCA" was selected for our study, as the 6 variables (BT consumption per capita and 5 key health indicators) are quantitative variables and this analysis was calibrated to study potential correlations.

Correlation circle

The correlation circle shows a projection of the initial variables in a dimensional space represented by axes F1 and F2. Variables are presented as vectors from the centre. When two vectors are close to the correlation circle, they can be: i) close to each other, meaning a positive correlation ii) orthogonal from each other, meaning that they are not correlated iii) on the opposite side from the centre, meaning a significant negative correlation. When some vectors are close to the centre, this means that some information is carried on other axes, and that any interpretation might be hazardous. This can be confirmed by looking at another correlation circle constructed with axes F1 and F3 or with axes F2 and F3. The correlation circle is then used to identify the potential proximity with the 6 vectors and to assess their potential correlations. Should a vector representing the variable "BT consumption" be close to the correlation circle and point to a similar direction compared to any of the other 5 vectors representing health indicators, this would indicate a positive correlation between the two variables.

Linear correlation model

Once identified by the correlation circle, potential correlations between BT consumption and one or more health indicators can be described using key statistical parameters, such as the coefficient r^2 and the statistical significance p. Using a linear correlation model between BT consumption and one health indicator then determines the extent to which the values of these two variables are potentially "proportional" to each other (BT consumption increases or decreases with one specific disease prevalence). The linear model formula is: $\mathbf{y = ax + b}$ (y= health indicator; x = BT consumption; a and b are the model calculated coefficients). The structure of this formula suggests that the variables x and y are linearly related and thus proportional; that is, the correlation is high if it can be represented by a straight line (upwards or downwards slope). If so, this line will represent the linear model, also called a "regression line" or "least squares line" because the sum of the squared distances of all the data points from the line is the lowest possible. The coefficient r^2 (coefficient of determination) represents the proportion of common variations between the two variables and establishes the "strength" of the relationship. In order to evaluate the potential correlation between BT consumption and one specific health indicator, it is therefore important to know r^2, the statistical significance p of the correlation (calculated by a Fisher-Snedecor test) and the statistical significance of the difference to 0 of the coefficient "a" (Student's t-test).

Results

The normative PCA deployed on the database was composed of 300 fields representing 6 variables (5 health indicators and BT consumption) in 50 countries. After mathematical projections of this multidimensional table, the overall "quality" (percentage of original variance) of the final projection in 2 dimensions was 59% and 74% when projected in 3 dimensions, confirming that the best representation of the dataset should be in 3 dimensions described by axes F1, F2 and F3. The "BT consumption" variable provided a high contributed 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, Viet-Nam, Philippine and Israel, suggesting the absence of any correlation between BT consumption and health indicators in these countries.

The "correlation circle" (Figure 1) shows that the "BT consumption" vector was strictly opposite the "Diabetes prevalence" vector, establishing a strong statistical negative correlation. Vectors concerning the other key health indicators (infectious diseases, respiratory diseases, cancer and cardiovascular diseases) were represented with a large angle (close to orthogonal) compared to the BT vector, confirming poor statistical relationships between BT and these 4 Health indicators. 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.

The linear correlation model can be expressed as follows:

Diabetes prevalence = a * BT consumption + b

Based on 42 countries, the p value of the Fisher-Snedecor test was 0.01, which is highly significant, confirming the relevance of the linear model. The coefficient r^2 was equal to 0.501.

The coefficient a =-0.0171183 and a Student's t-test confirmed that this coefficient was significantly different from 0 (p=0.001) with a 5% confidence interval between [-0.007; -0.027]. The negativity of the coefficient "a" means that when BT consumption increases, diabetes prevalence decreases, confirming a negative correlation (Figure 2).

Then linear correlation model can be represented by the following formula:

Diabetes prevalence = - 0.0171183 * BT consumption + 6173.64

Discussion

This study establishes, for the first time, a linear statistical relationship between high BT consumption and low diabetes prevalence in the countries that formed the basis for this analysis. As in any database analysis, the very first limitation of this study is related to the quality of the data. WHO prevalence data were obtained from the WHS, which constitutes a convenient and official source of key morbidity indicators around the world. The general design of the WHS is based on population sampling organized in the 192 Member States of the United Nations using face-to-face or telephone interviews. As the survey questionnaire offers a menu of choices of modules for each country, and lets the country select the survey approach (Household face-to-face survey, Computer-Assisted Telephone Interview or Computer-Assisted Personal Interview), the quality of data collection can be expected to be heterogeneous around the world. Furthermore, diabetes diagnostic criteria can vary from country to country. On the other hand, any fixed survey design with fixed criteria would not be appropriate everywhere, for example in countries with low telephone network coverage when planning telephone interviews. Other approaches to estimate prevalence of diabetes in the world have been studies using literature and data extrapolations [1], confirming the growing burden of diabetes.

Another important concern is the interpretation of the established statistical relationship between BT consumption and diabetes prevalence. Using advanced data mining techniques, we tested the potential statistical relationship between BT consumption and 5 health indicators, without any *a priori* assumptions in relation to any of these health indicators. We observed that, among the 5 health indicators, only the "prevalence of diabetes" indicator appeared to have a strong statistical relationship with BT consumption. The relevance and mechanism of this relationship then needs to be discussed. As tea is the most widely used ancient hot beverage in the world, the simple act of putting tea leaves into hot water has provided ancient societies with a tasty beverage with potential medicinal benefits. Two principal varieties of the species are used: the small-leaved Chinese variety (*C. sinensis sinensis*), also used for green tea and white tea, and the large-leaved Assamese variety (*C. sinensis assamica*), which has been traditionally only used for BT. Ancient Chinese civilizations realised that using a special fermentation process, tea leaves would become darker allowing them to be stored for longer periods of time. During this fermentation process, in which green tea oxidises to form black tea, caffeine tends to remain constant, while the types of flavonoids present in the tea differ. Green tea contains simple flavonoids called catechins, whilst BT contains complex flavonoids called theaflavins and thearubigins, which could be the chemical entities responsible for a number of potential health benefits. These tea types were called black tea because of the change in colour of the leaves as a result of this fermentation process. Numerous *in vitro* and *in vivo* studies have demonstrates the health benefits of green tea, mainly in cancer, cardiovascular disease, chronic inflammations or cognitive functions [14-22]. However, large-scale clinical dose-effect studies are still missing and it is difficult to interpret the clinically significance of results derived from some biological studies. Considerably fewer studies have been conducted on BT, mostly investigating its antioxidant properties [23, 24], and cardiovascular effects [25, 26]. Anti-diabetes properties of BT are suggested in some very specific studies such as a change in pancreatic function in streptozotocin-induced glucose-intolerant rats [27, 28], but also in some human studies together with other hot beverages [29-32]. Relatively recent interest in BT may be explained by the fact that BT is historically the type of tea most widely consumed in Western countries, probably due to its good storage properties, promoting active trade with tea-producing countries in Asia. Although there has recently been a renewed interest in green tea in industrialized countries due to its popular health benefits, BT represents over ninety percent of all tea sold in the West.

The obesity epidemic in many countries has stimulated interest in food components that may support weight management. Whereas many laboratory and physiological studies have demonstrated the potential effectiveness of BT for the prevention of obesity [28, 29, 33, 34], the underlying mechanisms remain unclear. The results of human intervention studies are mixed [35] and the role of caffeine has been suggested but not clearly established [34, 36]. Neyestani *et al* [33] found that regular daily intake of BT improves oxidative stress biomarkers and decreases serum C-reactive protein levels in type 2 diabetic patients. Histological studies on pancreas cells published by Manikandan *et al* [28] concluded that the BT extract contributes to regeneration of damaged pancreas cells and protects pancreatic beta cells by its antioxidant action. Nonetheless, the role of environment, dietary and lifestyle practices is fundamental when comparing health indicators around the world. Psaltopoulou *et al* [37] confirmed that low-glycaemic index dietary patterns reduced both fasting blood glucose and glycated proteins independently of carbohydrate consumption. Diets rich in whole-grain, cereal high-fibre products, and non-oil-seed pulses would also be beneficial. As vitamins and minerals play an important role in glucose metabolism, understanding the impact of potential vitamin and mineral deficiencies across cultures is also relevant to better organization of prevention and management of type 2 diabetes [38, 39]. An observational study based on nearly 37,000 middle-aged Chinese reported a 14% reduction in the risk of developing type 2 diabetes by drinking one or more cups of tea per day [40]. This was confirmed by two meta-analyses published by Huxley *et al* [34] and Jing *et al* [41]. Flavonoids are believed to support normal glucose metabolism via anti-inflammatory effects and increased insulin activity [42, 43]. Various studies, especially in Asian populations, confirm that flavonoids present in tea could reduce fat absorption in the gut, would promote fat oxidation in tissues and would increase energy expenditure [44]. An observational study of 4,300 Dutch adults found that flavonoid intake was highest in women who gained the least weight over a 14-year period [45]. Furthermore, as physical activity with or without diet contributes to a healthier lifestyle, this important factor must be considered when comparing health indicators between industrialized and emerging countries. Given rapid population growth, increased urbanization, and adverse lifestyle changes, the obesity/type 2 diabetes epidemic in resource-poor nations was predicted in the 1990s and has now been fully confirmed [46], underlying the importance of a better understanding of predictive and potentially protective factors.

The number of factors contributing to the growth of diabetes and obesity in the world confirm that "correlation does not imply causality ", and that a significant linear correlation between BT consumption and diabetes prevalence does not imply that low BT consumption could *cause* diabetes. If one factor is established as causing another, then the two factors are most certainly correlated. However, the opposite cannot be concluded. Thus, a correlation can only indicate a potential direct or indirect possible cause, which then needs to be further investigated. This paradigm and the connotations of causality may be the most important considerations affecting biostatistics in major epidemiological study designs [47]. A well known example of epidemiological cause-and-effect misinterpretations is the correlation that was established between hormone replacement therapy and a lower incidence of coronary heart disease. This association has been more recently explained by the fact that women taking hormone replacement therapy were more likely to come from higher socio-economic levels, which could explain the lower incidence of coronary heart disease [48]. Establishing causality is one of the most difficult challenges in public health. For instance, in clinical research, randomized controlled clinical trials are performed to establish potential significant differences between two groups. However, establishing a difference is not a demonstration of causality. Another example is case-control studies, which compare individuals with a specific disease ("cases") with a group of individuals without the disease ("controls"). An association between the hypothesized exposure and the disease studied would be reflected by a higher proportion in exposed cases, but this cannot constitute a real demonstration of causality. A potential causality can only be established with the convergence of interdisciplinary scientific evidence (biological, physiological, epidemiological, etc.) and reasonable explanations based on longitudinal studies. In our study, biological, physiological and some epidemiological studies can be considered to provide evidence linking BT consumption of BT and the prevalence of diabetes. 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.

Beyond the causality issue, a frequent criticism of using data-mining was based on the confusion between *data-mining* and *data-dredging* techniques. While a data-mining approach is based on searching for combinations of variables that might show potential correlations, data-dredging (also called "data-fishing") can generate misleading results [49]. When a number of hypotheses are tested, it is expected that some will falsely appear to be statistically significant, since every database can contain potential random correlations. A robust data mining approach must therefore always be based on a clear research strategy and a limited number of relevant meaningful assumptions. In our assessment, we used a systematic data mining approach to test potential correlations between 6 selected variables (BT consumption and 5 key health indicators). PCA was used to describe and structure the dataset before testing any correlations. In our study, only one linear correlation model was constructed between BT consumption and diabetes prevalence, based on the most relevant association suggested by the PCA. This consistent approach is quite different from screening numerous cross-regression analyses between all variables of one particular dataset. The data-mining approach can be considered to be a "radar tracking system", allowing detection, tracking and classification of potential "targets" in the framework of a particular environment. This is particularly useful when exploring complex databases, as data-mining can identify original statistical evidence, which would never be discovered by means of classical statistical techniques. As an example, the significant progress in genomics would not have been possible without the use of data-mining techniques [50]. Despite the data collection homogeneity issue inherent to large cross-country comparisons, we believe that this multidimensional approach could provide valuable additional scientific information, which is why our findings establishing a strong correlation between high BT consumption and low diabetes prevalence in these countries should be considered as a contribution to existing biological, physiological and epidemiological studies conducted on tea consumption, diabetes and obesity. These results should support further causality research regarding the health benefits of BT consumption on type 2 diabetes prevalence in the world.

References

1. Shaw JE, Sicree RA, Zimmet PZ. Global estimates of the prevalence of diabetes for 2010 and 2030. *Diabetes Research and Clinical Pratice.* 2010(87):4-14.

2. Emerging Risk Factors Collaboration, Sarwar N, Gao P, et al. Diabetes mellitus, fasting blood glucose concentration, and risk of vascular disease: a collaborative meta-analysis of 102 prospective studies. *Lancet.* 2010;375(9733):2215-2222.

3. Cheung N, Mitchell P, Wong TY. Diabetic retinopathy. *Lancet.* 2010;376(9735):124-136.

4. Danaei G, Finucane MM, Lu Y, et al. National, regional, and global trends in fasting plasma glucose and diabetes prevalence since 1980: systematic analysis of health examination surveys and epidemiological studies with 370 country-years and 2•7 million participants. *Lancet.* 2011;10.1016/S0140-6736(11):60679-X.

5. Ramachandran A MR, Snehalatha C. Diabetes in Asia. *Lancet.* 2010;375(9712):408-418.

6. Zimmet P, Alberti KG, Shaw J. Global and societal implications of the diabetes epidemic. *Nature Reviews.* 2001;414(6865):782-787.

7. Euromonitor. Hot Drinks: trade sources. 2010;www.euromonitor.com.

8. WHO. Global Health Survey. 2009;http://www.who.int/healthinfo/survey/en/.

9. Naqa I, Deasy J, Mu Y, et al. Datamining approaches for modeling tumor control probability. *Acta Oncol.* 2010;49(8):1363-1373.

10. Zhang F, Chen J. Data mining methods in omics-based biomarker discovery. *Methods Mol Biol.* 2011(719):511-526.

11. Wei CK SS, Yang MC. Application of Data Mining on the Development of a Disease Distribution Map of Screened Community Residents of Taipei County in Taiwan. *J Med Syst.* 2011(Feb 25).

12. Harpaz R, Haerian K, Chase H, Friedman C. Statistical Mining of Potential Drug Interaction Adverse Effects in FDA's Spontaneous Reporting System. *AMIA Annu Symp Proc.* 2010(Nov 13):281-285.

13. Briand S, Beresniak A, Nguyen T, et al. Assessment of yellow fever epidemic risk: an original multi-criteria modeling approach. *PLoS Negl Trop Dis.* 2009;3(7):e483.

14. Kang H, Rha S, Oh K, Nam C. Green tea consumption and stomach cancer risk: a meta-analysis. *Epidemiol Health.* 2010;32:e2010001.

15. Iwasaki M, Inoue M, Sasazuki S, et al. Green tea drinking and subsequent risk of breast cancer in a population to based cohort of Japanese women. *Breast Cancer Res.* 2010;12(5):R88.

16. Lee A, Liang W, Hirayama F, Binns C. Association between green tea consumption and lung cancer risk. *J Prev Med Public Health.* 2010;43(4):366-367.

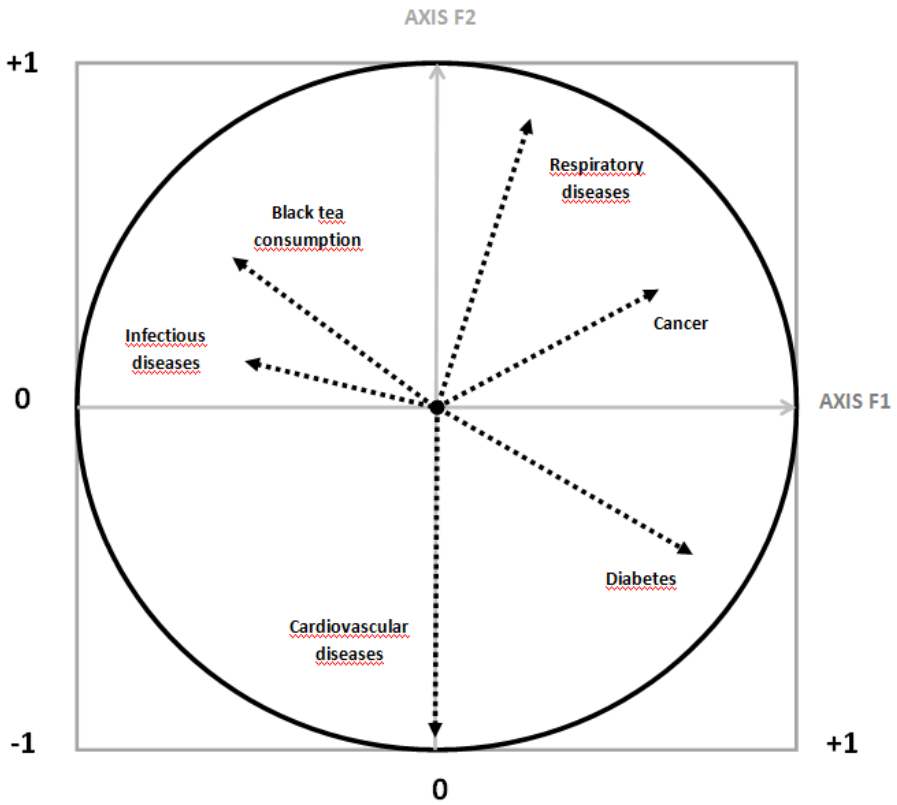
17. Moore RJ JK, Minihane AM. Green tea (Camellia sinensis) catechins and vascular function. *Br J Nutr.* 2009;102(12):1790-1802.

18. Feng L, Gwee X, Kua E, Ng T. Cognitive function and tea consumption in community dwelling older Chinese in Singapore. *J Nutr Health Aging.* 2010;14(6):433-438.

19. de Mejia E, Ramirez-Mares M, Puangpraphant S. Bioactive components of tea: cancer, inflammation and behavior. *Brain Behav Immun.* 2009;23(6):721-731.
20. Béliveau R, Gingras D. Green tea: prevention and treatment of cancer by nutraceuticals. *Lancet.* 2004;364(9439):1021-1022.
21. Walsh G. Tea and heart disease. *Lancet.* 1997;349(9053):735.
22. Ras R, Zock P, Draijer R. Tea Consumption Enhances Endothelial-Dependent Vasodilation; a Meta-Analysis. *PLoS ONE* 2011;6(3):e16974.
23. Pękal A, Dróżdż P, Biesaga M, Pyrzynska K. Evaluation of the antioxidant properties of fruit and flavoured black teas. *Eur J Nutr.* 2011;Mar 1.
24. Adhikary B, Yadav S, Roy K, Bandyopadhyay S, Chattopadhyay S. Black tea and theaflavins assist healing of indomethacin-induced gastric ulceration in mice by antioxidative action. *Evid Based Complement Alternat Med.* 2011(Sep 29): pii: 546560.
25. Bahorun T, Luximon-Ramma A, Gunness T, et al. Black tea reduces uric acid and C-reactive protein levels in humans susceptible to cardiovascular diseases. *Toxicology.* 2010;278(1):68-74.
26. Tokudome S, Nahomi I, Goto C, Tokudome Y, Moore M. Black tea and cardiovascular disease. *Int J Epidemiol.* 2005;34(2):482-483.
27. Dias T, Bronze MR, Houghton PJ, Mota-Filipe H, Paulo A. The flavonoid-rich fraction of Coreopsis tinctoria promotes glucose tolerance regain through pancreatic function recovery in streptozotocin-induced glucose-intolerant rats. *J Ethnopharmacol.* 2010;132(2):483-490.
28. Manikandan R SR, Thiagarajan R, Sivakumar MR, Meiyalagan V, Arumugam M. Effect of black tea on histological and immunohistochemical changes in pancreatic tissues of normal and streptozotocin-induced diabetic mice (Mus musculus). *Microsc Res Tech.* 2009;72(10):723-726.
29. Oba S NC, Nakamura K, Fujii K, Kawachi T, Takatsuka N, Shimizu H. Consumption of coffee, green tea, oolong tea, black tea, chocolate snacks and the caffeine content in relation to risk of diabetes in Japanese men and women. *Br J Nutr.* 2010;103(3):453-459.
30. Isogawa A, Noda M, Takahashi Y, Kadowaki T, Tsugane S. Coffee consumption and risk of type 2 diabetes mellitus. *Lancet.* 2003;361(9358):703-704.
31. Yoshioka K, Kogure A, Yoshida T, Yoshikawa T. Coffee consumption and risk of type 2 diabetes mellitus. *Lancet.* 2003;361(9358):703.
32. Reunanen A, Heliövaara M, Aho K. Coffee consumption and risk of type 2 diabetes mellitus. *Lancet.* 2003;361(9358):702-703.
33. Neyestani T, Shariatzade N, Kalayi A, et al. Regular daily intake of black tea improves oxidative stress biomarkers and decreases serum C-reactive protein levels in type 2 diabetic patients. *Ann Nutr Metab.* 2010;57(1):40-49.
34. Huxley R, Lee C, Barzi F, et al. Coffee, decaffeinated coffee, and tea consumption in relation to incident type 2 diabetes mellitus: a systematic review with meta-analysis. *Arch Intern Med.* 2009;169(22):2053-2063.
35. Hayashino Y, Fukuhara S, Okamura T, Tanaka T, Ueshima H, Group. H-OR. High oolong tea consumption predicts future risk of diabetes among Japanese male workers: a prospective cohort study. *Diabet Med.* 2011(Jan 18).
36. Goto A, Song Y, Chen B, Manson J, Buring J, Liu S. Coffee and caffeine consumption in relation to sex hormone-binding globulin and risk of type 2 diabetes in postmenopausal women. *Diabetes.* 2011;60(1):269-275.
37. Psaltopouloum T, Ilias I, Alevizaki M. The role of diet and lifestyle in primary, secondary, and tertiary diabetes prevention: a review of meta-analyses. *Rev Diabet Stud.* . 2010;7(1):26-35.
38. Martini L, Catania A, Ferreira S. Role of vitamins and minerals in prevention and management of type 2 diabetes mellitus. *Nutr Rev.* 2010;68(6):341-354.
39. Suksomboon N, Poolsup N, Sinprasert S. Effects of vitamin E supplementation on glycaemic control in type 2 diabetes: systematic review of randomized controlled trials. *J Clin Pharm Ther.* 2011;36(1):53-63.
40. Odegaard A, Pereira M, Koh W, Arakawa K, Lee H, Yu M. Coffee, tea and incident type 2 diabetes:
the Singapore Chinese Health Study. *American Journal of Clinical Nutrition.* 2008;88(4):979-985.

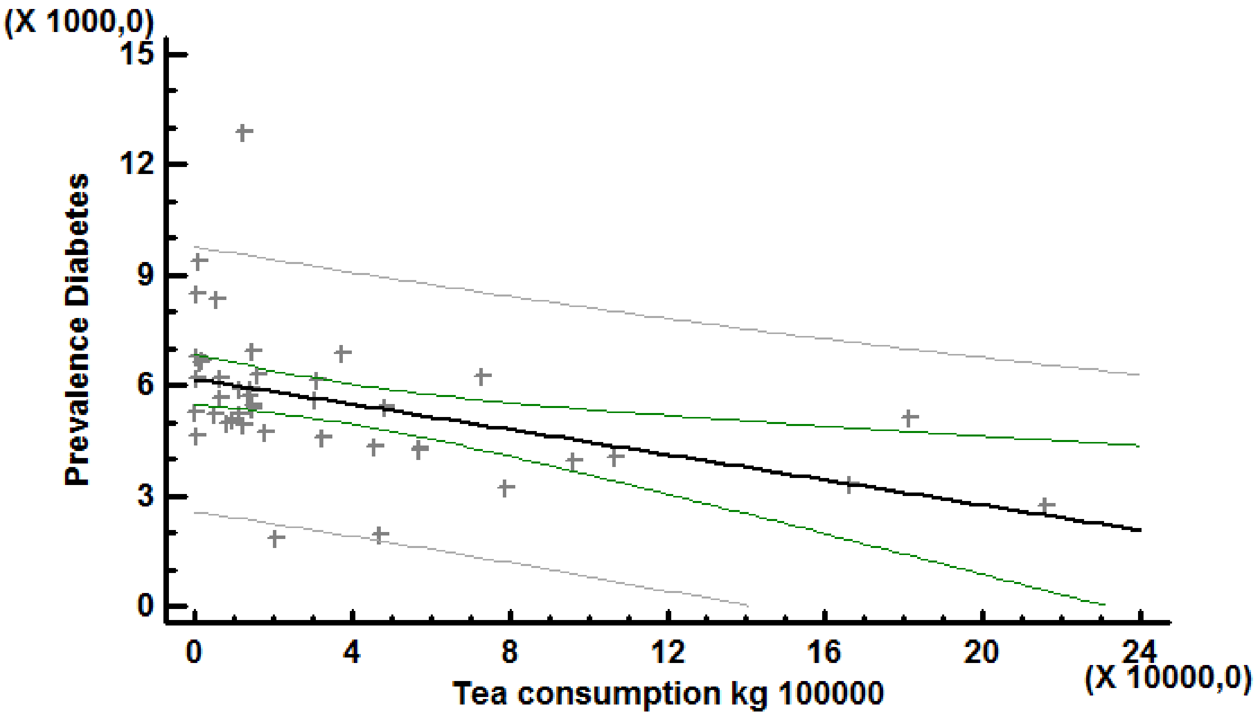
41. Jing Y, Han G, Hu Y, Bi Y, Li L, Zhu D. Tea consumption and risk of type 2 diabetes: a metaanalysis
of cohort studies. *J Gen Intern Med.* . 2009;24(5):557-562.
42. Nicolle E, Souard F, Faure P, Boumendjel A. Flavonoids as promising lead compounds in type 2 diabetes mellitus: molecules of interest and structure-activity relationship. *Curr Med Chem.* 2011;18(17):2661-2672.
43. Miyata Y, Tanaka H, Shimada A, et al. Regulation of adipocytokine secretion and adipocyte hypertrophy by polymethoxyflavonoids, nobiletin and tangeretin. *Life Sci.* . 2011;88(13-14):613-618.
44. MS. W-P. Green tea catechins, caffeine and body-weight regulation. *Physiol Behav.* 2010;100(1):42-46.
45. Hughes L, Arts I, Ambergen T, et al. Higher dietary flavone, flavonol, and catechin intakes are associated with less of an increase in BMI over time in women: a longitudinal analysis from the Netherlands Cohort Study. *Am J Clin Nutr.* 2008;88(5):1341-1352.
46. Nour N. Obesity in resource-poor nations. *Rev Obstet Gynecol.* . 2010;3(4):180-184.
47. Ortega Calvo M, Román Torres P, Lapetra Peralta J. Epistemology as health research propedeutics. *Gac Sanit.* 2011.
48. Lawlor D, Davey Smith G, Ebrahim S. Commentary: the hormone replacement-coronary heart disease conundrum: is this the death of observational epidemiology? *Int J Epidemiol.* 2004;33(3):464-467.
49. Lord S, Gebski V, Keech A. Multiple analyses in clinical trials: sound science or data dredging? *Med J Aust.* . 2004;18(18).
50. Lee J, Williams P, Cheon S. Data mining in genomics. *Clin Lab Med.* 2008;28(1):145-166.

Figure 1: Two dimensional Correlation circle of the 5 Health indicators and BT consumption*



**In this two dimensional representation, the "infectious disease" vector seems to be close to the BT vector, but is actually represented by a large angle in 3 dimensions, confirming the poor meaningful correlations between the "infectious diseases" and "BT consumption" variables.*

Figure 2: Linear correlation model between black tea consumption (kg per 100,000 inhabitants) and diabetes prevalence (cases per 100,000)



Data Mining Approach to Assess Statistical Relationships Between Black Tea Consumption and Key Health Indicators in the World

Journal:	*BMJ Open*
Manuscript ID:	bmjopen-2011-000648.R1
Article Type:	Research
Date Submitted by the Author:	02-Feb-2012
Complete List of Authors:	BERESNIAK, ARIEL; Data Mining International, Duru, Gerard; Data Mining International, Research Berger, Genevieve; Unilever, R&D Bremond-Gignac, Dominique; Amiens University Hospital, INSERM UMRS 968
Primary Subject Heading:	Epidemiology
Secondary Subject Heading:	Diabetes and endocrinology, Nutrition and metabolism, Public health
Keywords:	DIABETES & ENDOCRINOLOGY, NUTRITION & DIETETICS, EPIDEMIOLOGY

Data Mining Approach to Assess Statistical Relationships Between Black Tea Consumption and Key Health Indicators in the World

Ariel Beresniak, MD, MPH, PhD, Data Mining International, Geneva, Switzerland
Gerard Duru, PhD, Data Mining International; Geneva, Switzerland
Genevieve Berger, MD, PhD, Unilever, London, UK
Dominique Bremond-Gignac, MD, PhD, Amiens University Hospital, Amiens, INSERM UMRS 968, Paris VI University, France

Corresponding author: The corresponding author has the right to grant on behalf of all authors and does grant on behalf of all authors, an exclusive licence on a worldwide basis to the BMJ Publishing Group Ltd to permit this article to be published in BMJ Open and any other BMJPGL products and sublicences such use and to exploit all subsidiary rights, as set out in our licence.

Ariel Beresniak, MD, MPH, PhD
Data Mining International
Route de l'Aeroport, 29-31
CP221
CH-1215 Geneva 15
Switzerland
Phone: + 41 22 799 34 00 Fax: + 41 22 788 38 50
aberesniak@datamining-international.com

Funding statements: This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Competing interest statement: 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.

Acknowledgements: The authors thank Professor Paul Zimmet, Director Emeritus, Baker IDI Heart and Diabetes Institute, for reviewing the manuscript and making comments.

Data sharing statements: This study used epidemiological data from WHO World Health Surveys.

Individual contributions: The work presented here was carried out in collaboration between all authors. A.B. conceived the study aims and design. G.B. contributed to the data collection. A.B. and G.D performed the analysis. A.B, G.D and D.B. interpreted the results. A.B drafted the manuscript. All authors have contributed to, seen and approved the manuscript.

Abstract

Objectives: The objective of this study is to investigate potential statistical relationships between Black Tea consumption and five key health indicators (respiratory diseases, infectious diseases, cardiovascular diseases, diabetes, cancer).

Design: Data Mining analyses of Black Tea consumption data and World Health Survey data (WHO)

Setting: Comparative ecological study in 50 countries

Participants: No individual participants

Primary and secondary outcome measures: Primary outcomes measures were Black Tea consumption (in Kg/year per inhabitant) and prevalence of health indicators (100'000 inhabitant)

Results: Principal component analysis established a very high contribution of the black tea consumption parameter on the 3rd axis (81%). 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. A linear correlation model then confirmed a significant statistical correlation between high black tea consumption and low diabetes prevalence.

Conclusion: This innovative study establishes, for the first time, a linear statistical correlation between high black tea consumption and low diabetes prevalence in the world. These results are consistent with biological and physiological studies conducted on the potential effect of black tea on diabetes. Further epidemiological research and randomized studies are necessary to investigate the causality.

Trial registration: No trial (no registration)

Article summary

Article focus:

This study investigates potential statistical relationships between Black Tea consumption and a selection of key health indicators in 50 countries.

Key messages:

- A significant linear correlation was established between high black tea consumption and low diabetes prevalence.

- These results are consistent with biological and physiological studies conducted on the potential effect of black tea on diabetes and obesity.

- These results should support further causality research regarding the health benefits of BT consumption on type 2 diabetes prevalence in the world.

Strengths and limitations:

- These original study results are consistent with biological and physiological studies conducted on the potential effect of black tea on diabetes and obesity. We believe that this multidimensional approach provides valuable additional scientific information, as our findings, establishing a strong correlation between high BT consumption and low diabetes prevalence, can be considered to provide a contribution to existing biological, physiological and epidemiological studies conducted on tea consumption, diabetes and obesity.

- Diabetes prevalence data were obtained from the World Health Survey implemented by the World Health Organization, which constitutes an official source of key morbidity indicators around the world. However, the quality of data collection can be expected to be heterogeneous around the world and diabetes diagnostic criteria can vary from country to country.

- Another important concern is the interpretation of the established statistical relationship between BT consumption and diabetes prevalence. The numerous factors contributing to the growth of diabetes and obesity throughout the world confirm that "correlation does not imply causality", and that a significant linear correlation between BT consumption and diabetes prevalence does not imply that low BT consumption could *cause* diabetes. A correlation can only indicate a potential direct or indirect cause, which then needs to be further investigated.

- A frequent criticism of the use f data mining is based on the confusion between *data mining* and *data dredging* techniques. While a data mining approach is based on searching for combinations of variables that might show potential correlations, data dredging can generate misleading results. When a number of hypotheses are tested, it is expected that some will falsely appear to be statistically significant, since every database can contain potential random correlations. A robust data mining approach must therefore always be based on a clear research strategy and a limited number of relevant meaningful assumptions.

Background

An almost 6-fold increase in the number of people with diabetes has been observed over the last few decades. The International Diabetes Federation (IDF) reports that the number of people with diabetes will escalate from 285 million to 438 million between 2010 and 2030 [1] and the number of persons with IGT will increase from 344 to 472 million. By 2030, there will be over 900 million people worldwide with diabetes or at high risk of diabetes. Diabetes confers about a two-fold excess risk for a wide range of vascular diseases [2]. Furthermore, diabetic retinopathy is a common and specific microvascular complication of diabetes, and remains the leading cause of preventable blindness in working-aged people [3]. With one of the highest prevalences of all human diseases, diabetes is now a global epidemic with devastating health, social and economic consequences [4]. In certain ethnic groups, such as Asian populations, diabetes develops at a younger age than in Caucasian populations. Several distinctive features are apparent in the pathogenic factors for diabetes and their thresholds in Asian populations [5]. In conjunction with genetic susceptibility, type 2 diabetes is brought on by environmental and behavioural factors such as a sedentary lifestyle, overly rich nutrition and obesity and results in a huge economic burden [6]. It could therefore be interesting to investigate some key dietary habits in relation with lifestyle and health effects at a global level. For example, the positive health effects of black tea (BT) have been observed for centuries [7, 8].

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.

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.

Material and Method

Data sources

BT consumption data were derived from a specific international trade survey compiling sales data conducted in 2009 by Euromonitor International, an independent agency specialized in market research [9]. Consumption data are derived from black tea international trading registries, used by black tea importers to adapt international orders to local sales. Yearly consumption data expressed in kilograms per capita were available for the following 50 countries: Argentina, Australia, Austria, Belgium, Brazil, Bulgaria, Canada, Chile, China, Colombia, Czech Republic, Denmark, Egypt, Finland, France, Germany, Greece, Hungary, India, Indonesia, Ireland, Israel, Italy, Japan, Malaysia, Mexico, Morocco, Netherlands, New

Zealand, Norway, Philippines, Poland, Portugal, Romania, Russia, Saudi Arabia, Singapore, Slovakia, South Africa, South Korea, Spain, Sweden, Switzerland, Thailand, Turkey, Ukraine, United Kingdom, USA, Venezuela, Vietnam (Figure 1). Highest BT consumptions (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), as the Chinese population drinks 30 times more green tea (0.036 kg per inhabitant) than black tea.

Epidemiological data were derived from a specific analysis of the World Health Survey (WHS) conducted by the World Health Organization (WHO). Each year, the WHS compiles comprehensive baseline information on the health of populations and health system outcomes [10]. Using the 2009 dataset, five key health indicators were selected in 50 countries in both men and women for all age groups: prevalence of respiratory diseases, prevalence of infectious diseases (tuberculosis and HIV), prevalence of cancer, prevalence of cardiovascular diseases and prevalence of diabetes (Figure 2).

Methods

Data analyses were based on a systematic data mining approach. Data mining (sometimes called data or knowledge discovery) is generally defined as the process of analysing data from different perspectives and summarising these data into meaningful information. This approach is useful to analyse data derived from different dimensions or perspectives and to detect potential relationships between variables. Technically, data mining consists of discovering specific correlations or patterns in large relational databases. Data mining combines methods from statistics and artificial intelligence with database management and is considered to be an increasingly important tool. It is currently used in a wide range of scientific applications in health [11-14].

In this study, the data mining approach used 3 phases: firstly, a "calibrated principal component analyses" (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.

Normative principal component analysis (PCA)

PCA is a mathematical procedure that uses mathematical projections to convert a set of n possibly correlated variables representing n dimensions into a smaller number of dimensions called "principal components" classically represented in 2 or 3 axes F1, F2, F3. The projections use orthogonal transformations defined in such a way that the first principal component (first axis) has the highest possible variance in order to synthesize most of the initial information. The main objective of PCA is to reduce the dimensionality of the data set. PCA is often presented as a technique of factor analysis for quantitative variables. Multiple Correspondence Analysis (MCA) is another type of factor analysis for quantitative, qualitative and categorical variables and is useful to conduct multi-criteria analyses such as multi-criteria risk assessment [15]. A "normative PCA" was selected for our study, as the 6 variables (BT consumption per capita and 5 key health indicators) are quantitative variables and this analysis was calibrated to study potential correlations.

Correlation circle

The correlation circle shows a projection of the initial variables in a dimensional space represented by axes F1 and F2 [16]. Variables are presented as vectors from the centre. When two vectors are close to the correlation circle, they can be: i) close to each other, meaning a positive correlation ii) orthogonal from each other, meaning that they are not correlated iii) on the opposite side from the centre, meaning a significant negative correlation. When some vectors are close to the centre, this means that some information is carried on other axes, and that any interpretation might be hazardous. This can be confirmed by looking at another correlation circle constructed with axes F1 and F3 or with axes F2 and F3. The correlation circle is then used to identify the potential proximity with the 6 vectors and to assess their potential correlations. Should a vector representing the variable "BT consumption" be close to the correlation circle and point to a similar direction compared to any of the other 5 vectors representing health indicators, this would indicate a positive correlation between the two variables.

Linear correlation model

Once identified by the correlation circle, potential correlations between BT consumption and one or more health indicators can be described using key statistical parameters, such as the coefficient r^2 and the statistical significance p. Using a linear correlation model between BT consumption and one health indicator then determines the extent to which the values of these two variables are potentially "proportional" to each other (BT consumption increases or decreases with one specific disease prevalence). The linear model formula is: $\mathbf{y = ax + b}$ (y= health indicator; x = BT consumption; a and b are the model calculated coefficients). The structure of this formula suggests that the variables x and y are linearly related and thus proportional; that is, the correlation is high if it can be represented by a straight line (upwards or downwards slope). If so, this line will represent the linear model, also called a "regression line" or "least squares line" because the sum of the squared distances of all the data points from the line is the lowest possible. The coefficient r^2 (coefficient of determination) represents the proportion of common variations between the two variables and establishes the "strength" of the relationship. In order to evaluate the potential correlation between BT consumption and one specific health indicator, it is therefore important to know r^2, the statistical significance p of the correlation (calculated by a Fisher-Snedecor test) and the statistical significance of the difference to 0 of the coefficient "a" (Student's t-test).

Results

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.

The "correlation circle" (Figure 3) shows that the "BT consumption" vector was strictly opposite the "Diabetes prevalence" vector, establishing a strong statistical negative correlation. Vectors concerning the other key health indicators (infectious diseases, respiratory diseases, cancer and cardiovascular diseases) were represented with a large angle (close to orthogonal) compared to the BT vector, confirming poor statistical relationships between BT and these 4 Health indicators. 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. 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.

The linear correlation model can be expressed as follows:

Diabetes prevalence = a * BT consumption + b

Based on 42 countries, the p value of the Fisher-Snedecor test was 0.01, which is highly significant, confirming the relevance of the linear model. The coefficient r^2 was equal to 0.501.

The coefficient a =-0.0171183 and a Student's t-test confirmed that this coefficient was significantly different from 0 (p=0.001) with a 5% confidence interval between [-0.007; -0.027]. The negativity of the coefficient "a" means that when BT consumption increases, diabetes prevalence decreases, confirming a negative correlation (42).

Then linear correlation model can be represented by the following formula and is presented in Figure 4::

Diabetes prevalence = - 0.0171183 * BT consumption + 6173.64

Discussion

This study establishes, for the first time, a linear statistical relationship between high BT consumption and low diabetes prevalence in the countries that formed the basis for this analysis. As in any database analysis, the very first limitation of this study is related to the quality of the data. WHO prevalence data were obtained from the WHS, which constitutes a convenient and official source of key morbidity indicators around the world. The general design of the WHS is based on population sampling organized in the 192 Member States of the United Nations using face-to-face or telephone interviews. As the survey questionnaire offers a menu of choices of modules for each country, and lets the country select the survey approach (Household face-to-face survey, Computer-Assisted Telephone Interview or Computer-Assisted Personal Interview), the quality of data collection can be expected to be heterogeneous around the world.

Furthermore, 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. On the other hand, any fixed survey design with fixed criteria would not be appropriate everywhere, for example in countries with low telephone network coverage when planning telephone interviews. Other approaches to estimate prevalence of diabetes in the world have been studies using literature and data extrapolations [1], confirming the growing burden of diabetes.

Another important concern is the interpretation of the established statistical relationship between BT consumption and diabetes prevalence. Using advanced data mining techniques, we tested the potential statistical relationship between BT consumption and 5 health indicators, without any *a priori* assumptions in relation to any of these health indicators. We observed that, among the 5 health indicators, only the "prevalence of diabetes" indicator appeared to have a strong statistical relationship with BT consumption. 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. As this is the most widely used ancient hot beverage in the world, the simple act of putting tea leaves into hot water has provided ancient societies with a tasty beverage associated with the observation of certain medicinal benefits. Two principal varieties of the species are used: the small-leaved Chinese variety (*C. sinensis sinensis*), also used for green tea and white tea, and the large-leaved Assamese variety (*C. sinensis assamica*), which has been traditionally used only for BT. Ancient Chinese civilizations realised that using a special fermentation process, tea leaves would become darker allowing them to be stored for longer periods of time. During this fermentation process, in which green tea oxidises to form black tea, caffeine tends to remain constant, while the types of flavonoids present in the tea differ. Green tea contains simple flavonoids called catechins, whilst BT contains complex flavonoids called theaflavins and thearubigins, which could be the chemical entities responsible for a number of potential health benefits. These tea types were called black tea because of the change in colour of the leaves as a result of this fermentation process. Numerous *in vitro* and *in vivo* studies have demonstrated the health benefits of green tea, mainly in cancer, cardiovascular disease, chronic inflammation or cognitive functions [17-25]. However, large-scale clinical dose-effect studies are still missing and it is difficult to interpret the clinical significance of results derived from some biological studies. Considerably fewer studies have been conducted on BT, mostly investigating its antioxidant properties [26, 27], and cardiovascular effects [28, 29]. Anti-diabetes properties of BT have been suggested by several very specific studies, such as a change in pancreatic function in streptozotocin-induced glucose-intolerant rats [30, 31], but also in some human studies also investigating other hot beverages [32-35]. The relatively recent interest in BT may be explained by the fact that BT is historically the type of tea most widely consumed in Western countries, probably due to its good storage properties, promoting active trade with tea-producing countries in Asia. Although there has recently been a renewed interest in green tea in industrialized countries due to its popular health benefits, BT represents over ninety percent of all tea sold in the West.

The type 2 diabetes epidemic in many countries has stimulated interest in food components that may support weight management. 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 [36].

Although many laboratory studies have observed physiological effects of BT on glucose metabolism [31, 32, 37, 38], the underlying mechanisms remain unclear. The results of human intervention studies are mixed [39] and the role of caffeine has been suggested but not clearly established [38, 40]. Neyestani *et al* [37] found that regular daily intake of BT improves oxidative stress biomarkers and decreases serum C-reactive protein levels in type 2 diabetic patients. Histological studies on pancreas cells published by Manikandan *et al* [31] concluded that the BT extract contributes to regeneration of damaged pancreas cells and protects pancreatic beta cells by its antioxidant action. Nonetheless, the role of environment, dietary and lifestyle practices is fundamental when comparing health indicators around the world. Psaltopoulou *et al* [41] confirmed that low-glycaemic index dietary patterns reduced both fasting blood glucose and glycated proteins independently of carbohydrate consumption. Diets rich in whole-grain, cereal high-fibre products, and non-oil-seed pulses would also be beneficial. As vitamins and minerals play an important role in glucose metabolism, understanding the impact of potential vitamin and mineral deficiencies across cultures is also relevant to better organization of prevention and management of type 2 diabetes [42, 43]. An observational study based on nearly 37,000 middle-aged Chinese reported a 14% reduction in the risk of developing type 2 diabetes by drinking one or more cups of tea per day [44]. This was confirmed by two meta-analyses published by Huxley *et al* [38] and Jing *et al* [45]. Flavonoids are believed to support normal glucose metabolism via anti-inflammatory effects and increased insulin activity [46, 47]. 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 [48]. An observational study of 4,300 Dutch adults found that flavonoid intake was highest in women who gained the least weight over a 14-year period [49]. Furthermore, as physical activity with or without diet contributes to a healthier lifestyle, this important factor must be considered when comparing health indicators between industrialized and emerging countries. Given rapid population growth, increased urbanization, and adverse lifestyle changes, the obesity/type 2 diabetes epidemic in resource-poor nations was predicted in the 1990s and has now been fully confirmed [50], underlying the importance of a better understanding of predictive and potentially protective factors.

The number of factors contributing to the growth of diabetes and obesity in the world confirm that "correlation does not imply causality ", and that a significant linear correlation between BT consumption and diabetes prevalence does not imply that low BT consumption could *cause* diabetes. If one factor is established as causing another, then the two factors are most certainly correlated. However, the opposite cannot be concluded. Thus, a correlation can only indicate a potential direct or indirect possible cause, which then needs to be further investigated. This paradigm and the connotations of causality may be the most important considerations affecting biostatistics in major epidemiological study designs [51]. A well known example of epidemiological cause-and-effect misinterpretations is the correlation that was established between hormone replacement therapy and a lower incidence of coronary heart disease. This association has been more recently explained by the fact that women taking hormone replacement therapy were more likely to come from higher socio-economic levels, which could explain the lower incidence of coronary heart disease [52]. Establishing causality is one of the most difficult challenges in public health. For instance, in clinical research, randomized controlled clinical trials are performed to establish potential significant differences between two groups. However, establishing a difference is not a demonstration of causality. Another example is case-control studies, which compare individuals with a specific disease ("cases") with a group of individuals without the disease ("controls"). An association between the hypothesized exposure and the disease studied would be reflected by a higher proportion in exposed cases, but this cannot constitute a real demonstration of causality. A potential causality can only be established with the convergence of interdisciplinary scientific evidence (biological, physiological, epidemiological, etc.) and reasonable explanations based on longitudinal studies.

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 [26, 30, 31, 37-39, 46, 47]. 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.

Beyond the causality issue, a frequent criticism of using data mining was based on the confusion between *data mining* and *data dredging* techniques. While a data mining approach is based on searching for combinations of variables that might show potential correlations, data-dredging (also called "data fishing") can generate misleading results [53]. When a number of hypotheses are tested, it is expected that some will falsely appear to be statistically significant, since every database can contain potential random correlations. A robust data mining approach must therefore always be based on a clear research strategy and a limited number of relevant meaningful assumptions. In our assessment, we used a systematic data mining approach to test potential correlations between 6 selected variables (BT consumption and 5 key health indicators). PCA was used to describe and structure the dataset before testing any correlations. In our study, only one linear correlation model was constructed between BT consumption and diabetes prevalence, based on the most relevant association suggested by the PCA. This consistent approach is quite different from screening numerous cross-regression analyses between all variables of one particular dataset. The data mining approach can be considered to be a "radar tracking system", allowing detection, tracking and classification of potential "targets" in the framework of a particular environment. This is particularly useful when exploring complex databases, as data mining can identify original statistical evidence, which would never be discovered by means of classical statistical techniques. As an example, the significant progress in genomics would not have been possible without the use of data mining techniques [54]. Despite the data collection homogeneity issue inherent to large cross-country comparisons, we believe that this multidimensional approach could provide valuable additional scientific information, which is why our findings establishing a strong correlation between high BT consumption and low diabetes prevalence in these countries should be considered as a contribution to existing biological, physiological and epidemiological studies conducted on tea consumption, diabetes and obesity. These results should support further causality research regarding the health benefits of BT consumption on type 2 diabetes prevalence in the world.

References

1. Shaw JE, Sicree RA, Zimmet PZ. Global estimates of the prevalence of diabetes for 2010 and 2030. *Diabetes Research and Clinical Pratice.* 2010(87):4-14.

2. Emerging Risk Factors Collaboration, Sarwar N, Gao P, et al. Diabetes mellitus, fasting blood glucose concentration, and risk of vascular disease: a collaborative meta-analysis of 102 prospective studies. *Lancet.* 2010;375(9733):2215-2222.

3. Cheung N, Mitchell P, Wong TY. Diabetic retinopathy. *Lancet.* 2010;376(9735):124-136.

4. Danaei G, Finucane MM, Lu Y, et al. National, regional, and global trends in fasting plasma glucose and diabetes prevalence since 1980: systematic analysis of health examination surveys and epidemiological studies with 370 country-years and 2•7 million participants. *Lancet.* 2011;10.1016/S0140-6736(11):60679-X.

5. Ramachandran A MR, Snehalatha C. Diabetes in Asia. *Lancet.* 2010;375(9712):408-418.

6. Zimmet P, Alberti KG, Shaw J. Global and societal implications of the diabetes epidemic. *Nature Reviews.* 2001;414(6865):782-787.

7. 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.

8. 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.

9. Euromonitor. Hot Drinks: trade sources. 2010;www.euromonitor.com.

10. WHO. Global Health Survey. 2009;http://www.who.int/healthinfo/survey/en/.

11. Naqa I, Deasy J, Mu Y, et al. Datamining approaches for modeling tumor control probability. *Acta Oncol.* 2010;49(8):1363-1373.

12. Zhang F, Chen J. Data mining methods in omics-based biomarker discovery. *Methods Mol Biol.* 2011(719):511-526.

13. Wei CK SS, Yang MC. Application of Data Mining on the Development of a Disease Distribution Map of Screened Community Residents of Taipei County in Taiwan. *J Med Syst.* 2011(Feb 25).

14. Harpaz R, Haerian K, Chase H, Friedman C. Statistical Mining of Potential Drug Interaction Adverse Effects in FDA's Spontaneous Reporting System. *AMIA Annu Symp Proc.* 2010(Nov 13):281-285.

15. Briand S, Beresniak A, Nguyen T, et al. Assessment of yellow fever epidemic risk: an original multi-criteria modeling approach. *PLoS Negl Trop Dis.* 2009;3(7):e483.

16. Everitt B, Dunn G. Applied Multivariate data analysis. *Lavoisier Publisher.* 2001, 2d ed.:320p

17. Kang H, Rha S, Oh K, Nam C. Green tea consumption and stomach cancer risk: a meta-analysis. *Epidemiol Health.* 2010;32:e2010001.

18. Iwasaki M, Inoue M, Sasazuki S, et al. Green tea drinking and subsequent risk of breast cancer in a population to based cohort of Japanese women. *Breast Cancer Res.* 2010;12(5):R88.

19. Lee A, Liang W, Hirayama F, Binns C. Association between green tea consumption and lung cancer risk. *J Prev Med Public Health.* 2010;43(4):366-367.

20. Moore RJ JK, Minihane AM. Green tea (Camellia sinensis) catechins and vascular function. *Br J Nutr.* 2009;102(12):1790-1802.
21. Feng L, Gwee X, Kua E, Ng T. Cognitive function and tea consumption in community dwelling older Chinese in Singapore. *J Nutr Health Aging.* 2010;14(6):433-438.
22. de Mejia E, Ramirez-Mares M, Puangpraphant S. Bioactive components of tea: cancer, inflammation and behavior. *Brain Behav Immun.* 2009;23(6):721-731.
23. Béliveau R, Gingras D. Green tea: prevention and treatment of cancer by nutraceuticals. *Lancet.* 2004;364(9439):1021-1022.
24. Walsh G. Tea and heart disease. *Lancet.* 1997;349(9053):735.
25. Ras R, Zock P, Draijer R. Tea Consumption Enhances Endothelial-Dependent Vasodilation; a Meta-Analysis. *PLoS ONE* 2011;6(3):e16974.
26. Pękal A, Dróżdż P, Biesaga M, Pyrzynska K. Evaluation of the antioxidant properties of fruit and flavoured black teas. *Eur J Nutr.* 2011;Mar 1.
27. Adhikary B, Yadav S, Roy K, Bandyopadhyay S, Chattopadhyay S. Black tea and theaflavins assist healing of indomethacin-induced gastric ulceration in mice by antioxidative action. *Evid Based Complement Alternat Med.* 2011(Sep 29): pii: 546560.
28. Bahorun T, Luximon-Ramma A, Gunness TK, et al. Black tea reduces uric acid and C-reactive protein levels in humans susceptible to cardiovascular diseases. *Toxicology.* 2010;278(1):68-74.
29. Tokudome S, Nahomi I, Goto C, Tokudome Y, Moore M. Black tea and cardiovascular disease. *Int J Epidemiol.* 2005;34(2):482-483.
30. Dias T, Bronze MR, Houghton PJ, Mota-Filipe H, Paulo A. The flavonoid-rich fraction of Coreopsis tinctoria promotes glucose tolerance regain through pancreatic function recovery in streptozotocin-induced glucose-intolerant rats. *J Ethnopharmacol.* 2010;132(2):483-490.
31. Manikandan R SR, Thiagarajan R, Sivakumar MR, Meiyalagan V, Arumugam M. Effect of black tea on histological and immunohistochemical changes in pancreatic tissues of normal and streptozotocin-induced diabetic mice (Mus musculus). *Microsc Res Tech.* 2009;72(10):723-726.
32. Oba S NC, Nakamura K, Fujii K, Kawachi T, Takatsuka N, Shimizu H. Consumption of coffee, green tea, oolong tea, black tea, chocolate snacks and the caffeine content in relation to risk of diabetes in Japanese men and women. *Br J Nutr.* 2010;103(3):453-459.
33. Isogawa A, Noda M, Takahashi Y, Kadowaki T, Tsugane S. Coffee consumption and risk of type 2 diabetes mellitus. *Lancet.* 2003;361(9358):703-704.
34. Yoshioka K, Kogure A, Yoshida T, Yoshikawa T. Coffee consumption and risk of type 2 diabetes mellitus. *Lancet.* 2003;361(9358):703.
35. Reunanen A, Heliövaara M, Aho K. Coffee consumption and risk of type 2 diabetes mellitus. *Lancet.* 2003;361(9358):702-703.
36. Ang YG, Wu XC, Toh MP, Chia KS, Heng BH. Progression Rate of newly diagnosed Impaired Fasting Glycemia to Type 2 Diabetes Mellitus: a study using the National Healthcare Group Diabetes Registry in Singapore. *J Diabetes.* 2011;Nov 7.
37. Neyestani T, Shariatzade N, Kalayi A, et al. Regular daily intake of black tea improves oxidative stress biomarkers and decreases serum C-reactive protein levels in type 2 diabetic patients. *Ann Nutr Metab.* 2010;57(1):40-49.
38. Huxley R, Lee C, Barzi F, et al. Coffee, decaffeinated coffee, and tea consumption in relation to incident type 2 diabetes mellitus: a systematic review with meta-analysis. *Arch Intern Med.* 2009;169(22):2053-2063.
39. Hayashino Y, Fukuhara S, Okamura T, Tanaka T, Ueshima H, Group. H-OR. High oolong tea consumption predicts future risk of diabetes among Japanese male workers: a prospective cohort study. *Diabet Med.* 2011(Jan 18).
40. Goto A, Song Y, Chen B, Manson J, Buring J, Liu S. Coffee and caffeine consumption in relation to sex hormone-binding globulin and risk of type 2 diabetes in postmenopausal women. *Diabetes.* 2011;60(1):269-275.
41. Psaltopouloum T, Ilias I, Alevizaki M. The role of diet and lifestyle in primary, secondary, and tertiary diabetes prevention: a review of meta-analyses. *Rev Diabet Stud.* . 2010;7(1):26-35.
42. Martini L, Catania A, Ferreira S. Role of vitamins and minerals in prevention and management of type 2 diabetes mellitus. *Nutr Rev.* 2010;68(6):341-354.

43. Suksomboon N, Poolsup N, Sinprasert S. Effects of vitamin E supplementation on glycaemic control in type 2 diabetes: systematic review of randomized controlled trials. *J Clin Pharm Ther.* 2011;36(1):53-63.

44. Odegaard A, Pereira M, Koh W, Arakawa K, Lee H, Yu M. Coffee, tea and incident type 2 diabetes: the Singapore Chinese Health Study. *American Journal of Clinical Nutrition.* 2008;88(4):979-985.

45. Jing Y, Han G, Hu Y, Bi Y, Li L, Zhu D. Tea consumption and risk of type 2 diabetes: a metaanalysis of cohort studies. *J Gen Intern Med.* . 2009;24(5):557-562.

46. Nicolle E, Souard F, Faure P, Boumendjel A. Flavonoids as promising lead compounds in type 2 diabetes mellitus: molecules of interest and structure-activity relationship. *Curr Med Chem.* 2011;18(17):2661-2672.

47. Miyata Y, Tanaka H, Shimada A, et al. Regulation of adipocytokine secretion and adipocyte hypertrophy by polymethoxyflavonoids, nobiletin and tangeretin. *Life Sci.* . 2011;88(13-14):613-618.

48. MS. W-P. Green tea catechins, caffeine and body-weight regulation. *Physiol Behav.* 2010;100(1):42-46.

49. Hughes L, Arts I, Ambergen T, et al. Higher dietary flavone, flavonol, and catechin intakes are associated with less of an increase in BMI over time in women: a longitudinal analysis from the Netherlands Cohort Study. *Am J Clin Nutr.* 2008;88(5):1341-1352.

50. Nour N. Obesity in resource-poor nations. *Rev Obstet Gynecol.* . 2010;3(4):180-184.

51. Ortega Calvo M, Román Torres P, Lapetra Peralta J. Epistemology as health research propedeutics. *Gac Sanit.* 2011.

52. Lawlor D, Davey Smith G, Ebrahim S. Commentary: the hormone replacement-coronary heart disease conundrum: is this the death of observational epidemiology? *Int J Epidemiol.* 2004;33(3):464-467.

53. Lord S, Gebski V, Keech A. Multiple analyses in clinical trials: sound science or data dredging? *Med J Aust.* . 2004;18(18).

54. Lee J, Williams P, Cheon S. Data mining in genomics. *Clin Lab Med.* 2008;28(1):145-166.

Figure 1: 2009 Black Tea consumption data in kg/year per inhabitant (source: Euromonitor)

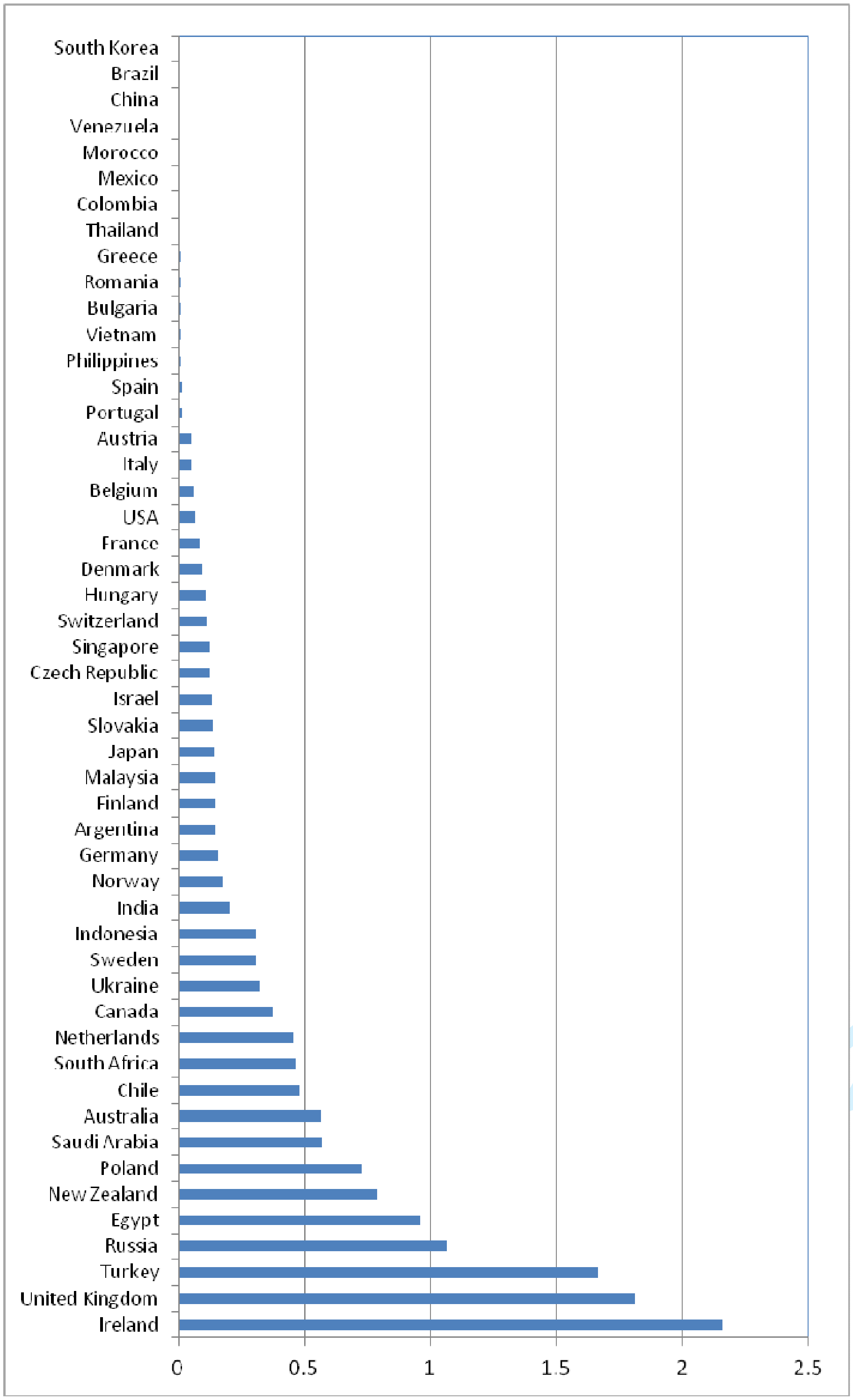


Figure 2: 2009 prevalence (per 100,000) of key health indicators (source: WHO)

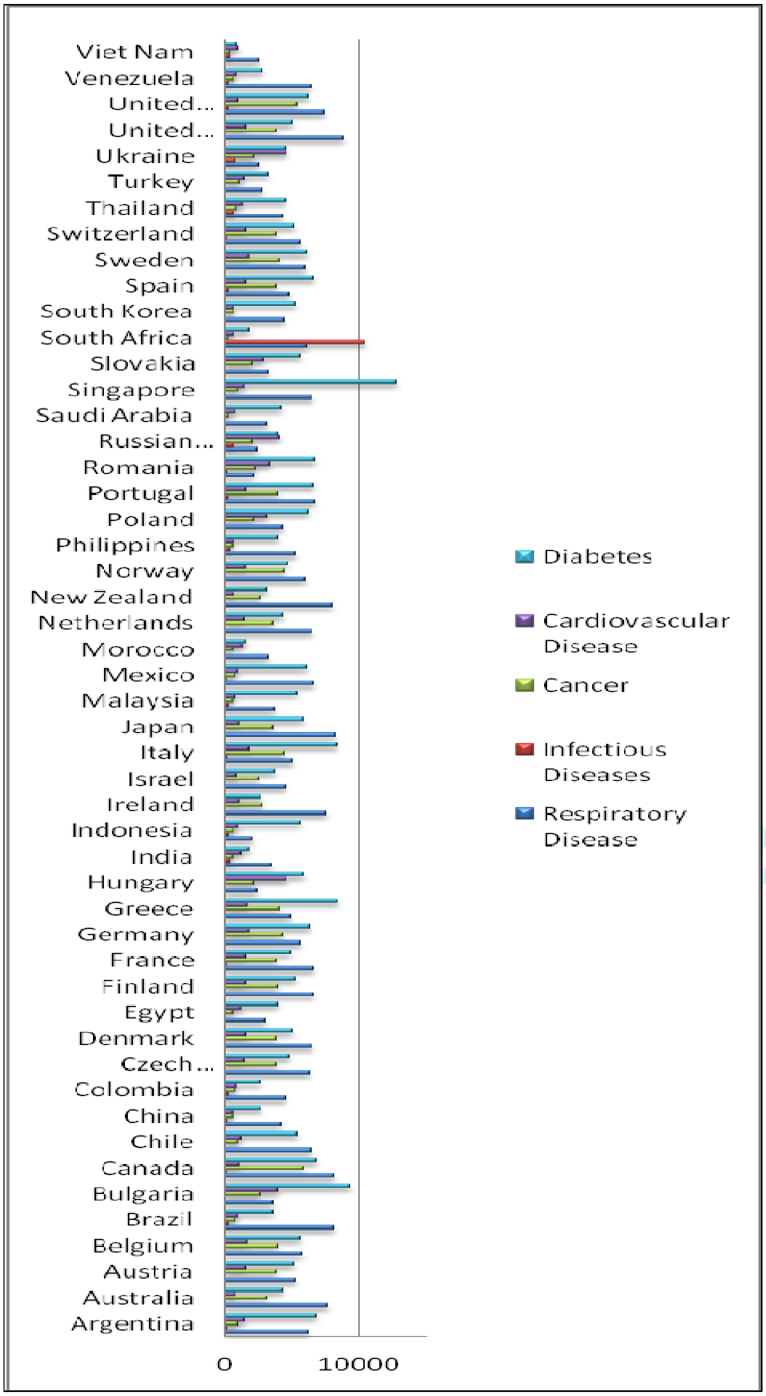
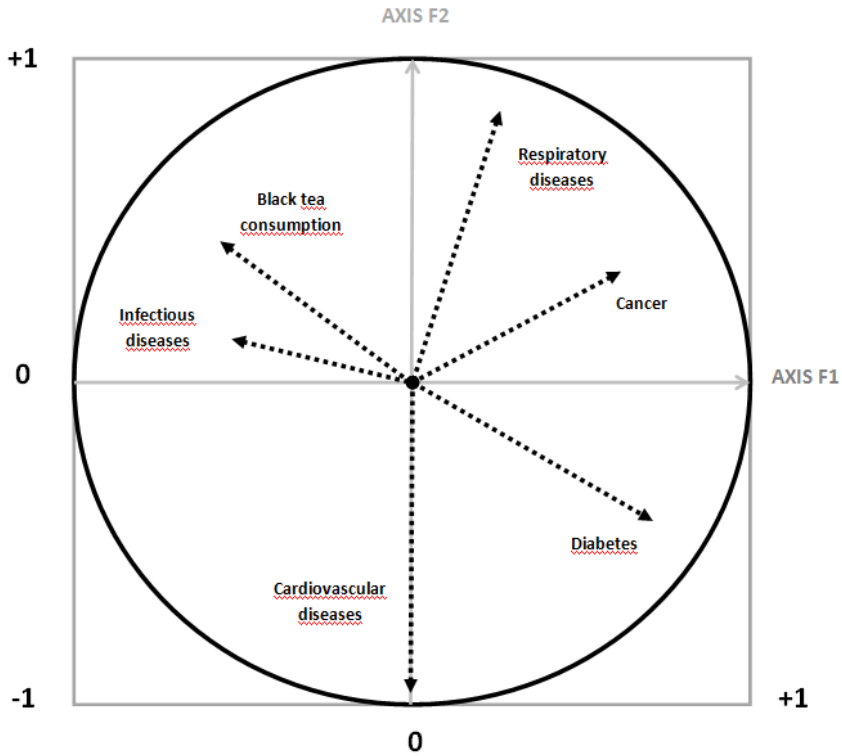
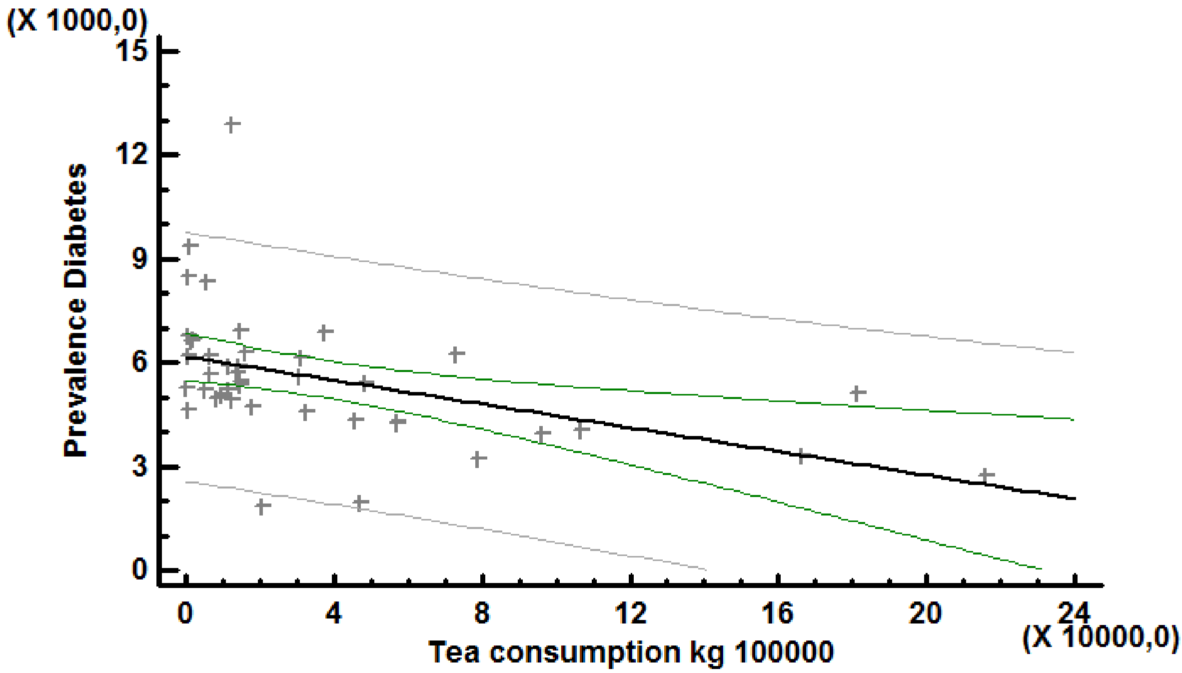


Figure 3: Two dimensional correlation circle of 5 health indicators and BT consumption*



**In this two-dimensional representation, the "infectious disease" vector seems to be close to the BT vector, but is actually represented by a large angle in 3 dimensions, confirming the poor meaningful correlations between the "infectious diseases" and "BT consumption" variables.*

Figure 4: Linear correlation model between black tea consumption (kg per 100,000 inhabitants) and diabetes prevalence (cases per 100,000)



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

Journal:	*BMJ Open*
Manuscript ID:	bmjopen-2011-000648.R2
Article Type:	Research
Date Submitted by the Author:	15-Aug-2012
Complete List of Authors:	BERESNIAK, ARIEL; Data Mining International, Duru, Gerard; Data Mining International, Research Berger, Genevieve; Unilever, R&D BREMOND-GIGNAC, Dominique; Amiens University Hospital, INSERM UMRS 968
Primary Subject Heading:	Epidemiology
Secondary Subject Heading:	Diabetes and endocrinology, Nutrition and metabolism, Public health
Keywords:	DIABETES & ENDOCRINOLOGY, NUTRITION & DIETETICS, EPIDEMIOLOGY

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

Ariel Beresniak, MD, MPH, PhD, Data Mining International, Geneva, Switzerland
Gerard Duru, PhD, Data Mining International; Geneva, Switzerland
Genevieve Berger, MD, PhD, Unilever, London, UK
Dominique Bremond-Gignac, MD, PhD, Amiens University Hospital, Amiens, INSERM UMRS 968, Paris VI University, France

Ariel Beresniak, MD, MPH, PhD
Data Mining International
Route de l'Aeroport, 29-31
CP221
CH-1215 Geneva 15
Switzerland
Phone: + 41 22 799 34 00 Fax: + 41 22 788 38 50
aberesniak@datamining-international.com

Key words: diabetes & endocrinology; Nutrition & dietetics; Epidemiology

Word count: 4325

Acknowledgements: The authors thank Professor Paul Zimmet, Director Emeritus, Baker IDI Heart and Diabetes Institute, for reviewing the manuscript and making comments.

Funding Statement : This research received no specific funding.

Contributorship Statement : The work presented here was carried out in collaboration between all authors. A.B. conceived the study aims and design. G.B. contributed to the data collection. A.B. and G.D performed the analysis. A.B, G.D and D.B. interpreted the results. A.B drafted the manuscript. All authors have contributed to, seen and approved the manuscript.

Data Sharing Statement : Authors agree to share the dataset used in this study (2009 Black Tea consumption data and 2009 health indicators), which will be made available by contacting the first author.

Conflict of Interests Statement : 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.

Abstract

Objective: The objective of this study was to investigate possible statistical relationships between black tea consumption and key health indicators in the world.

Methodology: This ecological study used a systematic data mining approach, which was carried out on black tea consumption data and five key health epidemiological indicators from the World Health Survey supervised by the World Health Organization: respiratory diseases, infectious diseases, cancer, cardiovascular diseases and diabetes. The methodological approach included 3 phases: firstly, a "calibrated principal component analysis" was used to segment the database composed of 6 variables (black tea consumption and 5 health indicators) into 3 dimensions; secondly, the 6 variables were represented as vectors in a projected "correlation circle" to study potential positive or negative correlations; lastly, a linear correlation model was tested on selected variables.

Results: Principal component analysis established a very high contribution of the black tea consumption parameter on the 3rd axis (81%). 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. A linear correlation model then confirmed a significant statistical correlation between high black tea consumption and low diabetes prevalence.

Conclusion: This innovative study establishes a linear statistical correlation between high black tea consumption and low diabetes prevalence in the world. These results are consistent with biological and physiological studies conducted on the effect of black tea on diabetes and confirm the results of a previous ecological study in Europe. Further epidemiological research and randomised studies are necessary to investigate the causality.

ARTICLE SUMMARY

Article focus:

This study investigates potential statistical relationships between Black Tea consumption and a selection of key health indicators in 50 countries.

Key messages:

- A significant linear correlation was established between high black tea consumption and low diabetes prevalence.
- These results are consistent with biological, physiological and ecological studies conducted on the potential effect of black tea on diabetes and obesity.
- These results should support further causality research regarding the health benefits of BT consumption on type 2 diabetes prevalence in the world.

Strengths and limitations:

- These original study results are consistent with previous biological, physiological and ecological studies conducted on the potential effect of black tea on diabetes and obesity. We believe that this multidimensional approach provides valuable additional scientific information at the global level, as our findings, establishing a strong correlation between high BT consumption and low diabetes prevalence, can be considered to provide a contribution to existing studies conducted on tea consumption, diabetes and obesity.

- Diabetes prevalence data were obtained from the World Health Survey implemented by the World Health Organization, which constitutes an official source of key morbidity indicators around the world. However, the quality of data collection can be expected to be heterogeneous around the world and diabetes diagnostic criteria can vary from country to country.
- Another important concern is the interpretation of the established statistical relationship between BT consumption and diabetes prevalence. The numerous factors contributing to the growth of diabetes and obesity throughout the world confirm that "correlation does not imply causality", and that a significant linear correlation between BT consumption and diabetes prevalence does not imply that low BT consumption could cause diabetes. A correlation can only indicate a potential direct or indirect cause, which then needs to be further investigated.
- A frequent criticism of the use of data mining is based on the confusion between data mining and data dredging techniques. While a data mining approach is based on searching for combinations of variables that might show potential correlations, data dredging can generate misleading results. When a number of hypotheses are tested, it is expected that some will falsely appear to be statistically significant, since every database can contain potential random correlations. A robust data mining approach must therefore always be based on a clear research strategy and a limited number of relevant meaningful assumptions.
-A classical 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. Our study on black tea does not comprise any potential logical fallacy, as it was not used as the basis for any individual assumptions.

Background

Various study designs have been used to assess the potential benefits of tea. As tea is the most widely used ancient hot beverage in the world, the simple act of putting tea leaves into hot water has provided ancient societies with a tasty beverage associated with the observation of certain medicinal benefits. Two principal varieties of the species are used: the small-leaved Chinese variety (*C. sinensis sinensis*), also used for green tea and white tea, and the large-leaved Assamese variety (*C. sinensis assamica*), which has been traditionally used only for black tea (BT). Ancient Chinese civilizations realised that by using a special fermentation process, tea leaves would become darker allowing them to be stored for longer periods of time. During this fermentation process, in which green tea oxidises to form black tea, caffeine tends to remain constant, while the types of flavonoids present in the tea differ. Green tea contains simple flavonoids called catechins, whilst BT contains complex flavonoids called theaflavins and thearubigins, which could be the chemical entities responsible for a number of potential health benefits. These tea types were called black tea because of the change in colour of the leaves as a result of this fermentation process. 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 [1-9]. 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 [10], protect against oxidative damage and apoptosis in human bronchial epithelial cells induced by tobacco 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 [11] and appear to exert a protective activity against *Vibrio cholerae*[12]. However, research into the potential beneficial effects of tea appears to be most active in the field of cardiovascular diseases, in view of the number 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 [13]. Moreover, dietary consumption of tea catechins would have beneficial effects on blood pressure and lipid parameters [14]. 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 randomised 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 [15]. Anti-diabetes properties of BT have been suggested by several very specific studies, such as a change in pancreatic function in streptozotocin-induced glucose-intolerant rats [16, 17], but also in some human studies investigating other hot beverages [18-21]. The relatively recent interest in BT may be explained by the fact that BT is historically the type of tea most widely consumed in Western countries, probably due to its good storage properties, promoting active trade with tea-producing countries in Asia. Although there has recently been a renewed interest in green tea in industrialized countries, BT represents over ninety percent of all tea sold in the West. Despite the number of publications investigating the effects of tea components and green tea in particular, large-scale clinical dose-effect studies are still lacking and it is difficult to interpret the clinical significance of results derived from some biological studies. Considerably fewer studies have been conducted specifically on BT, mostly investigating its antioxidant properties [22, 23], and cardiovascular effects [24, 25]. It could therefore be interesting to investigate some key dietary habits in relation with lifestyle and health effects at a global level, in view of the perceived positive health effects of black tea (BT), which have been described for centuries [26, 27]. Because of 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 using advanced exploratory statistical methods. 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.

Material and Method

Data sources

BT consumption data were derived from a specific international trade survey compiling sales data conducted in 2009 by Euromonitor International, an independent agency specialized in market research [28]. Consumption data are derived from black tea international trading registries, used by black tea importers to adapt international orders to local sales. Yearly consumption data expressed in kilograms per capita were available for the following 50 countries: Argentina, Australia, Austria, Belgium, Brazil, Bulgaria, Canada, Chile, China, Colombia, Czech Republic, Denmark, Egypt, Finland, France, Germany, Greece, Hungary, India, Indonesia, Ireland, Israel, Italy, Japan, Malaysia, Mexico, Morocco, Netherlands, New Zealand, Norway, Philippines, Poland, Portugal, Romania, Russia, Saudi Arabia, Singapore, Slovakia, South Africa, South Korea, Spain, Sweden, Switzerland, Thailand, Turkey, Ukraine, United Kingdom, USA, Venezuela, Vietnam (Figure 1). Highest BT consumptions (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), as the Chinese population drinks 30 times more green tea (0.036 kg per inhabitant) than black tea. Epidemiological data were derived from a specific extraction from the World Health Survey (WHS) conducted by the World Health Organization (WHO). Each year, the WHS compiles comprehensive baseline information on the health of populations and health system outcomes [29]. Using the 2009 dataset (sample presented in Table 1), five key health indicators were selected in 50 countries in both men and women for all age groups: prevalence of respiratory diseases, prevalence of infectious diseases (tuberculosis and HIV), prevalence of cancer, prevalence of cardiovascular diseases and prevalence of diabetes.

Methods

This ecological study used a data mining approach structured in 3 phases: 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 as the mathematical projection of the 6 dimensions defined by the 6 variables into 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.

Normative principal component analysis (PCA)

PCA is a mathematical procedure that uses mathematical projections to convert a set of *n* possibly correlated variables representing *n* dimensions into a smaller number of dimensions called "principal components" classically represented in 2 or 3 axes F1, F2, F3. The projections use orthogonal transformations defined in such a way that the first principal component (first axis) has the highest possible variance in order to synthesize most of the initial information. The main objective of PCA is to reduce the dimensionality of the dataset. PCA is often presented as a technique of factor analysis for quantitative variables. Multiple Correspondence Analysis (MCA) is another type of factor analysis for quantitative, qualitative and categorical variables and is useful to conduct multi-criteria analyses such as multi-criteria risk assessment [30]. A "normative PCA" was selected for our study, as the 6 variables (BT consumption per capita and 5 key health indicators) are quantitative variables and this analysis was calibrated to study potential correlations.

Correlation circle

The correlation circle shows a projection of the initial variables in a dimensional space which can be represented in two or three dimensions [31]. Variables are presented as vectors from the centre. When two vectors are close to the correlation circle, they can be: i) close to each other, meaning a positive correlation ii) orthogonal from each other, meaning that they are not correlated iii) on the opposite side from the centre, meaning a significant negative correlation. When some vectors are close to the centre, this means that some information is carried on other axes, and that any interpretation might be hazardous. The correlation circle is then used to identify the potential proximity with the 6 vectors and to assess their potential correlations. Should a vector representing the variable "BT consumption" be close to the correlation circle and point to a similar direction compared to any of the other 5 vectors representing health indicators, this would indicate a positive correlation between the two variables.

Linear correlation model

Once identified by the correlation circle, potential correlations between BT consumption and one or more health indicators can be described using key statistical parameters, such as the coefficient r^2 and the statistical significance *p*. Using a linear correlation model between BT consumption and one health indicator then determines the extent to which the values of these two variables are potentially "proportional" to each other (BT consumption increases or decreases with one specific disease prevalence). The linear model formula is: **y = ax + b** (y= health indicator; x = BT consumption; a and b are the model calculated coefficients). The structure of this formula suggests that the variables x and y are linearly related and thus proportional; that is, the correlation is high if it can be represented by a straight line (upwards or downwards slope). If so, this line will represent the linear model, also called a "regression line" or "least squares line" because the sum of the squared distances of all the data points from the line is the lowest possible. The coefficient r^2 (coefficient of determination) represents the proportion of common variations between the two variables and establishes the "strength" of the relationship. In order to evaluate the potential correlation between BT consumption and one specific health indicator, it is therefore important to know r^2, the statistical significance *p* of the correlation (calculated by a Fisher-Snedecor test) and the statistical significance of the difference to 0 of the coefficient "a" (Student's t-test).

Results

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.

The "correlation circle" (Figure 2) shows that the "BT consumption" vector was strictly opposite the "Diabetes prevalence" vector, establishing a strong statistical negative correlation. Vectors concerning the other key health indicators (infectious diseases, respiratory diseases, cancer and cardiovascular diseases) were represented with a large angle (close to orthogonal) compared to the BT vector, confirming poor statistical relationships between BT and these 4 Health indicators. Of particular interest was the interpretation of the "infectious disease" vector, which seemed to be close to the BT vector in a two dimensions projection , 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. 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.

Then linear correlation model with the format y= ax+b is represented by the following formula and is presented in Figure 3:

Diabetes prevalence = - 0.0171183 * BT consumption + 6173.64

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 be unknown (6,173 cases per 100,000 inhabitants). Based on 42 countries, the p value of the Fisher-Snedecor test was 0.003, which is highly significant, confirming the relevance of the linear model. The coefficient r^2 was equal to 0.199. Student's t-test confirmed that the slope coefficient (0.0171183) was significantly different from 0 (p=0.003) with a 5% confidence interval between [-0.028; -0.006]. The negativity of coefficient "a" indicates that diabetes prevalence decreases as BT consumption increases, confirming a negative correlation.

Discussion

Limitations

This study establishes an inverse linear statistical relationship between high BT consumption and diabetes prevalence in the world, and confirms the findings of the European ecological study establishing a similar relationship[32]. As in any database analysis, the very first limitation of this study is related to the quality of the data. WHO prevalence data were obtained from the WHS, which constitutes a convenient and official source of key morbidity indicators around the world. The general design of the WHS is based on population sampling organized in the 192 Member States of the United Nations using face-to-face or telephone interviews. As the survey questionnaire offers a menu of choices of modules for each country, and lets the country select the survey approach (Household face-to-face survey, Computer-Assisted Telephone Interview or Computer-Assisted Personal Interview), the quality of data collection can be expected to be heterogeneous around the world.

Furthermore, 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. On the other hand, any fixed survey design with fixed criteria would not be appropriate everywhere, for example in countries with low telephone network coverage when planning telephone interviews. Other approaches to estimate prevalence of diabetes in the world have been studies using literature and data extrapolations [33], confirming the growing burden of diabetes. Another important concern is the interpretation of the established statistical relationship between BT consumption and diabetes prevalence. Using a systematic data mining approach, we tested the potential statistical relationship between BT consumption and 5 health indicators, without any *a priori* assumptions in relation to any of these health indicators. We observed that, among the 5 health indicators, only the "prevalence of diabetes" indicator appeared to have a strong statistical relationship with BT consumption. This ecological approach considers the population as the unit of analysis rather than an individual, 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. A classical 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 the basis for any individual assumptions. However, when interesting and 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 studies was that published by Keys in 1980 [34] concerning the relationship with dietary habits and coronary heart disease in 7 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 [35]. For these reasons, 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.

Growing interest of food components that may support weight management and glucose metabolism

Our results confirm the recent 2012 publication from the InterAct Consortium which carried out an European ecological study confirming an inverse linear association between tea consumption and the incidence of type 2 diabetes in Europe [32]. The type 2 diabetes epidemic in many countries has stimulated interest in food components that may support weight management. An almost 6-fold increase in the number of people with diabetes has been observed over the last few decades. The International Diabetes Federation (IDF) reports that the number of people with diabetes will escalate from 285 million to 438 million between 2010 and 2030 [33] and the number of persons with IGT will increase from 344 to 472 million. By 2030, there will be over 900 million people worldwide with diabetes or at high risk of diabetes. Diabetes confers about a two-fold excess risk for a wide range of vascular diseases [36]. Furthermore, diabetic retinopathy is a common and specific microvascular complication of diabetes, and remains the leading cause of preventable blindness in working-aged people [37]. With one of the highest prevalences of all human diseases, diabetes is now a global epidemic with devastating health, social and economic consequences [38]. In certain ethnic groups, such as Asian populations, diabetes develops at a younger age than in Caucasian populations. Several distinctive features are apparent in the pathogenic factors for diabetes and their thresholds in Asian populations [39]. In conjunction with genetic susceptibility, type 2 diabetes is brought on by environmental and behavioural factors such as a sedentary lifestyle, overly rich nutrition and obesity and results in a huge economic burden [40]. According to WHS 2009 data, Singapore is the country with the highest diabetes prevalence with 12,876 cases per 100,000 inhabitants, which is mainly observed in the Chinese community and is probably due to the intense urban lifestyle in Singapore [41]. Although many laboratory studies have observed physiological effects of BT on glucose metabolism [17, 18, 42, 43], the underlying mechanisms remain unclear. The results of human intervention studies are mixed [44] and the role of caffeine has been suggested but not clearly established [43, 45]. Neyestani *et al* [42] found that regular daily intake of BT improves oxidative stress biomarkers and decreases serum C-reactive protein levels in type 2 diabetic patients. Histological studies on pancreas cells published by Manikandan *et al* [17] concluded that the BT extract contributes to regeneration of damaged pancreas cells and protects pancreatic beta cells by its antioxidant action. Nonetheless, the role of environment, dietary and lifestyle practices is fundamental when comparing health indicators around the world. Psaltopoulou *et al* [46] confirmed that low-glycaemic index dietary patterns reduced both fasting blood glucose and glycated proteins independently of carbohydrate consumption. Diets rich in whole-grain, cereal high-fibre products, and non-oil-seed pulses would also be beneficial. As vitamins and minerals play an important role in glucose metabolism, understanding the impact of potential vitamin and mineral deficiencies across cultures is also relevant to better organization of prevention and management of type 2 diabetes [47, 48]. An observational study based on nearly 37,000 middle-aged Chinese reported a 14% reduction in the risk of developing type 2 diabetes by drinking one or more cups of tea per day [49]. This was confirmed by two meta-analyses published by Huxley *et al* [43] and Jing *et al* [50]. Flavonoids are believed to support normal glucose metabolism via anti-inflammatory effects and increased insulin activity [51, 52]. 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 [53]. An observational study of 4,300 Dutch adults found that flavonoid intake was highest in women who gained the least weight over a 14-year period [54]. Furthermore, as physical activity with or without diet contributes to a healthier lifestyle, this important factor must be considered when comparing health indicators between industrialized and emerging countries. Given rapid population growth, increased urbanization, and adverse lifestyle changes, the obesity/type 2 diabetes epidemic in resource-poor nations was predicted in the 1990s and has now been fully confirmed [55], underlying the importance of a better understanding of predictive and potentially protective factors.

Correlation and causality

The number of factors contributing to the growth of diabetes and obesity in the world confirms that "correlation does not imply causality ", and that a significant linear correlation between BT consumption and diabetes prevalence does not imply that low BT consumption could *cause* diabetes. If one factor is established as causing another, then the two factors are most certainly correlated. However, the opposite cannot be concluded. Thus, a correlation can only indicate a potential direct or indirect possible cause, which then needs to be further investigated. This paradigm and the connotations of causality may be the most important considerations affecting biostatistics, not only in ecological studies but also in major epidemiological study designs [56]. A well known example of epidemiological cause-and-effect misinterpretations is the correlation that was established between hormone replacement therapy and a lower incidence of coronary heart disease. This association has been more recently explained by the fact that women taking hormone replacement therapy were more likely to come from higher socio-economic levels, which could explain the lower incidence of coronary heart disease [57]. Establishing causality is one of the most difficult challenges in public health. For instance, in clinical research, randomised controlled clinical trials are performed to establish potential significant differences between two groups. However, establishing a difference is not a demonstration of causality. Another example is case-control studies, which compare individuals with a specific disease ("cases") with a group of individuals without the disease ("controls"). An association between the hypothesized exposure and the disease studied would be reflected by a higher proportion in exposed cases, but this cannot constitute a real demonstration of causality. A potential causality can only be established with the convergence of interdisciplinary scientific evidence (biological, physiological, epidemiological, etc.) and reasonable explanations based on longitudinal studies. In any case, ecological research can address important issues that cannot be easily addressed by other study designs. Ecological studies are frequently used when alternative study designs are not possible (eg, randomised control trials), such as when investigating the effect of geographical factors on disease incidence. Our research, like all ecological studies and most other epidemiological approaches,presents a number of limitations 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. Despite the fact that a number of biological, physiological and epidemiological field studies have provided evidence linking BT consumption and glucose metabolism [16, 17, 22, 42-44, 51, 52], a large-scale randomised controlled trial of tea consumption and diabetes risk would be useful to confirm these findings.

Data mining and data dredging

Beyond the causality issue, a frequent criticism of using data mining was based on the confusion between *data mining* and *data dredging* techniques. While a data mining approach is based on searching for combinations of variables that might show potential correlations, data-dredging (also called "data fishing") can generate misleading results [58]. When a number of hypotheses are tested, it is expected that some will falsely appear to be statistically significant, since every database can contain potential random correlations. A robust data mining approach must therefore always be based on a clear research strategy and a limited number of relevant meaningful assumptions. In our assessment, we used a systematic data mining approach to test potential correlations between 6 selected variables (BT consumption and 5 key health indicators). PCA was used to describe and structure the dataset before testing any correlations. In our study, only one linear correlation model was constructed between BT consumption and diabetes prevalence, based on the most relevant association suggested by the PCA. This consistent approach is quite different from screening numerous cross-regression analyses between all variables of one particular dataset. The data mining approach can be considered to be a "radar tracking system", allowing detection, tracking and classification of potential "targets" in the framework of a particular environment. This is particularly useful when exploring complex databases, as data mining can identify original statistical evidence, which would never be discovered by means of classical statistical techniques. As an example, the significant progress in genomics would not have been possible without the use of data mining techniques. Despite the data collection homogeneity issue inherent to large cross-country comparisons, we believe that this multidimensional approach can provide valuable additional scientific information, completing published biological, physiological and epidemiological studies conducted on tea consumption, diabetes and obesity. These results should support further causality research regarding the health benefits of BT consumption on type 2 diabetes prevalence in the world.

References

1. Kang H, Rha S, Oh K, et al. Green tea consumption and stomach cancer risk: a meta-analysis. *Epidemiol Health.* 2010;32:e2010001.
2. Iwasaki M, Inoue M, Sasazuki S, et al. Green tea drinking and subsequent risk of breast cancer in a population to based cohort of Japanese women. *Breast Cancer Res.* 2010;12(5):R88.
3. Lee A, Liang W, Hirayama F, et al. Association between green tea consumption and lung cancer risk. *J Prev Med Public Health.* 2010;43(4):366-367.
4. Moore RJ JK, Minihane AM. Green tea (Camellia sinensis) catechins and vascular function. *Br J Nutr.* 2009;102(12):1790-1802.
5. Feng L, Gwee X, Kua E, et al. Cognitive function and tea consumption in community dwelling older Chinese in Singapore. *J Nutr Health Aging.* 2010;14(6):433-438.
6. de Mejia E, Ramirez-Mares M, Puangpraphant S. Bioactive components of tea: cancer, inflammation and behavior. *Brain Behav Immun.* 2009;23(6):721-731.
7. Béliveau R, Gingras D. Green tea: prevention and treatment of cancer by nutraceuticals. *Lancet.* 2004;364(9439):1021-1022.
8. Walsh G. Tea and heart disease. *Lancet.* 1997;349(9053):735.
9. Ras R, Zock P, Draijer R. Tea Consumption Enhances Endothelial-Dependent Vasodilation; a Meta-Analysis. *PLoS ONE* 2011;6(3):e16974.
10. Liang OD, Kleibrink BE, Schuette-Nuetgen K, et al. Green tea epigallo-catechin-galleate ameliorates the development of obliterative airway disease. *Exp Lung Res.* 2011;37(7):435-444.
11. 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):123-132.
12. Toda M, Okubo S, Ikigai H, et al. The protective activity of tea catechins against experimental infection by Vibrio cholerae O1. *Microbiol Immunol.* 1992;36(9):999-1001.
13. Islam MA. Cardiovascular effects of green tea catechins: progress and promise. . *Recent Pat Cardiovasc Drug Discov.* 2012;7(2):88-99.
14. Hodgson JM, Puddey IB, Woodman RJ, et al. Effects of black tea on blood pressure: a randomized controlled trial. *Arch Intern Med.* 2012;172(2):186-188.
15. 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):1259-1270.
16. Dias T, Bronze MR, Houghton PJ, Mota-Filipe H, Paulo A. The flavonoid-rich fraction of Coreopsis tinctoria promotes glucose tolerance regain through pancreatic function recovery in streptozotocin-induced glucose-intolerant rats. *J Ethnopharmacol.* 2010;132(2):483-490.
17. Manikandan R SR, Thiagarajan R, Sivakumar MR, Meiyalagan V, Arumugam M. Effect of black tea on histological and immunohistochemical changes in pancreatic tissues of normal and streptozotocin-induced diabetic mice (Mus musculus). *Microsc Res Tech.* 2009;72(10):723-726.
18. Oba S NC, Nakamura K, Fujii K, et al. Consumption of coffee, green tea, oolong tea, black tea, chocolate snacks and the caffeine content in relation to risk of diabetes in Japanese men and women. *Br J Nutr.* 2010;103(3):453-459.
19. Isogawa A, Noda M, Takahashi Y, et al. Coffee consumption and risk of type 2 diabetes mellitus. *Lancet.* 2003;361(9358):703-704.
20. Yoshioka K, Kogure A, Yoshida T, et al. Coffee consumption and risk of type 2 diabetes mellitus. *Lancet.* 2003;361(9358):703.
21. Reunanen A, Heliövaara M, Aho K. Coffee consumption and risk of type 2 diabetes mellitus. *Lancet.* 2003;361(9358):702-703.
22. Pękal A, Dróżdż P, Biesaga M, et al. Evaluation of the antioxidant properties of fruit and flavoured black teas. *Eur J Nutr.* 2011;Mar 1.

23. Adhikary B, Yadav S, Roy K, Bandyopadhyay S, et al. Black tea and theaflavins assist healing of indomethacin-induced gastric ulceration in mice by antioxidative action. *Evid Based Complement Alternat Med.* 2011(Sep 29): pii: 546560.
24. Bahorun T, Luximon-Ramma A, et al. Black tea reduces uric acid and C-reactive protein levels in humans susceptible to cardiovascular diseases. *Toxicology.* 2010;278(1):68-74.
25. Tokudome S, Nahomi I, Goto C, et al. Black tea and cardiovascular disease. *Int J Epidemiol.* 2005;34(2):482-483.
26. 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.
27. 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.
28. Euromonitor. Hot Drinks: trade sources. 2010;www.euromonitor.com.
29. WHO. Global Health Survey. 2009;http://www.who.int/healthinfo/survey/en/.
30. Briand S, Beresniak A, Nguyen T, et al. Assessment of yellow fever epidemic risk: an original multi-criteria modeling approach. *PLoS Negl Trop Dis.* 2009;3(7):e483.
31. Everitt B, Dunn G. Applied Multivariate data analysis. *Lavoisier Publisher.* 2001, 2d ed.:320p

32. Consortium. TI. Tea Consumption and Incidence of Type 2 Diabetes in Europe: The EPIC-InterAct Case-Cohort Study. *PLoS One.* 2012(7):5.
33. Shaw JE, Sicree RA, Zimmet PZ. Global estimates of the prevalence of diabetes for 2010 and 2030. *Diabetes Research and Clinical Practice.* 2010(87):4-14.
34. Keys A. Seven Countries: A Multivariate Analysis of Death and Coronary Heart Disease. *Harvard University Press.* 1980.
35. Didkowska J, Manczuk M, McNeill A, et al. Lung cancer mortality at ages 35-54 in the European Union: ecological study of evolving tobacco epidemics. *BMJ.* 2005;331(7510):189-191.
36. Emerging Risk Factors Collaboration, Sarwar N, Gao P, et al. Diabetes mellitus, fasting blood glucose concentration, and risk of vascular disease: a collaborative meta-analysis of 102 prospective studies. *Lancet.* 2010;375(9733):2215-2222.
37. Cheung N, Mitchell P, Wong TY. Diabetic retinopathy. *Lancet.* 2010;376(9735):124-136.
38. Danaei G, Finucane MM, Lu Y, et al. National, regional, and global trends in fasting plasma glucose and diabetes prevalence since 1980: systematic analysis of health examination surveys and epidemiological studies with 370 country-years and 2•7 million participants. *Lancet.* 2011;10.1016/S0140-6736(11):60679-X.
39. Ramachandran A MR, Snehalatha C. Diabetes in Asia. *Lancet.* 2010;375(9712):408-418.
40. Zimmet P, Alberti KG, Shaw J. Global and societal implications of the diabetes epidemic. *Nature Reviews.* 2001;414(6865):782-787.
41. Ang YG, Wu XC, Toh MP, et al. Progression Rate of newly diagnosed Impaired Fasting Glycemia to Type 2 Diabetes Mellitus: a study using the National Healthcare Group Diabetes Registry in Singapore. *J Diabetes.* 2011;Nov 7.
42. Neyestani T, Shariatzade N, Kalayi A, et al. Regular daily intake of black tea improves oxidative stress biomarkers and decreases serum C-reactive protein levels in type 2 diabetic patients. *Ann Nutr Metab.* 2010;57(1):40-49.
43. Huxley R, Lee C, Barzi F, et al. Coffee, decaffeinated coffee, and tea consumption in relation to incident type 2 diabetes mellitus: a systematic review with meta-analysis. *Arch Intern Med.* 2009;169(22):2053-2063.
44. Hayashino Y, Fukuhara S, Okamura T, et al. High oolong tea consumption predicts future risk of diabetes among Japanese male workers: a prospective cohort study. *Diabet Med.* 2011(Jan 18).
45. Goto A, Song Y, Chen B, et al. Coffee and caffeine consumption in relation to sex hormone-binding globulin and risk of type 2 diabetes in postmenopausal women. *Diabetes.* 2011;60(1):269-275.
46. Psaltopouloum T, Ilias I, Alevizaki M. The role of diet and lifestyle in primary, secondary, and tertiary diabetes prevention: a review of meta-analyses. *Rev Diabet Stud.* . 2010;7(1):26-35.
47. Martini L, Catania A, Ferreira S. Role of vitamins and minerals in prevention and management of type 2 diabetes mellitus. *Nutr Rev.* 2010;68(6):341-354.

48. Suksomboon N, Poolsup N, Sinprasert S. Effects of vitamin E supplementation on glycaemic control in type 2 diabetes: systematic review of randomized controlled trials. *J Clin Pharm Ther.* 2011;36(1):53-63.

49. Odegaard A, Pereira M, Koh W, et al. Coffee, tea and incident type 2 diabetes: the Singapore Chinese Health Study. *American Journal of Clinical Nutrition.* 2008;88(4):979-985.

50. Jing Y, Han G, Hu Y, et al. Tea consumption and risk of type 2 diabetes: a metaanalysis of cohort studies. *J Gen Intern Med.* . 2009;24(5):557-562.

51. Nicolle E, Souard F, Faure P, Boumendjel A. Flavonoids as promising lead compounds in type 2 diabetes mellitus: molecules of interest and structure-activity relationship. *Curr Med Chem.* 2011;18(17):2661-2672.

52. Miyata Y, Tanaka H, Shimada A, et al. Regulation of adipocytokine secretion and adipocyte hypertrophy by polymethoxyflavonoids, nobiletin and tangeretin. *Life Sci.* . 2011;88(13-14):613-618.

53. MS. W-P. Green tea catechins, caffeine and body-weight regulation. *Physiol Behav.* 2010;100(1):42-46.

54. Hughes L, Arts I, Ambergen T, et al. Higher dietary flavone, flavonol, and catechin intakes are associated with less of an increase in BMI over time in women: a longitudinal analysis from the Netherlands Cohort Study. *Am J Clin Nutr.* 2008;88(5):1341-1352.

55. Nour N. Obesity in resource-poor nations. *Rev Obstet Gynecol.* . 2010;3(4):180-184.

56. Ortega Calvo M, Román Torres P, Lapetra Peralta J. Epistemology as health research propedeutics. *Gac Sanit.* 2011.

57. Lawlor D, Davey Smith G, Ebrahim S. Commentary: the hormone replacement-coronary heart disease conundrum: is this the death of observational epidemiology? *Int J Epidemiol.* 2004;33(3):464-467.

58. Lord S, Gebski V, Keech A. Multiple analyses in clinical trials: sound science or data dredging? *Med J Aust.* . 2004;18(18).

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

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

Formatted: Font: 14 pt, Bold

~~Data Mining Approach to Assess Statistical Relationships Between Black Tea Consumption and Key Health Indicators in the World~~

Ariel Beresniak, MD, MPH, PhD, Data Mining International, Geneva, Switzerland
Gerard Duru, PhD, Data Mining International; Geneva, Switzerland
Genevieve Berger, MD, PhD, Unilever, London, UK
Dominique Bremond-Gignac, MD, PhD, Amiens University Hospital, Amiens, INSERM UMRS 968, Paris VI University, France

Ariel Beresniak, MD, MPH, PhD
Data Mining International
Route de l'Aeroport, 29-31
CP221
CH-1215 Geneva 15
Switzerland
Phone: + 41 22 799 34 00 Fax: + 41 22 788 38 50
aberesniak@datamining-international.com

Funding statements: This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Competing interest statement: 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.

Acknowledgements: The authors thank Professor Paul Zimmet, Director Emeritus, Baker IDI Heart and Diabetes Institute, for reviewing the manuscript and making comments.

Data sharing statements: This study used epidemiological data from WHO World Health Surveys.

Individual contributions: The work presented here was carried out in collaboration between all authors. A.B. conceived the study aims and design. G.B. contributed to the data collection. A.B. and G.D performed the analysis. A.B, G.D and D.B. interpreted the results. A.B drafted the manuscript. All authors have contributed to, seen and approved the manuscript.

Abstract

Objective: The objective of this study was to investigate a possible correlation between black tea consumption and key health indicators in the world.

Methodology: A systematic data mining approach was carried out on black tea consumption data and five key health epidemiological indicators from the World Health Survey (WHO): respiratory diseases, infectious diseases, cancer, cardiovascular diseases and diabetes. The methodological approach included 3 phases: firstly, a "calibrated principal component analysis" was used to segment the database composed of 6 variables (black tea consumption and 5 health indicators) into 3 dimensions; secondly, the 6 variables were represented as vectors in a projected "correlation circle" to study potential positive or negative correlations; lastly, a linear correlation model was tested on selected variables.

Results: Principal component analysis established a very high contribution of the black tea consumption parameter on the 3rd axis (81%). 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. A linear correlation model then confirmed a significant statistical correlation between high black tea consumption and low diabetes prevalence.

Conclusion: This innovative study establishes, for the first time, a linear statistical correlation between high black tea consumption and low diabetes prevalence in the world. These results are consistent with biological and physiological studies conducted on the potential effect of black tea on diabetes and confirms the results of a previous ecological study in Europe. Further epidemiological research and randomis~~z~~ed studies are necessary to investigate the causality.

Article summary

Article focus:

This study investigates potential statistical relationships between Black Tea consumption and a selection of key health indicators in 50 countries.

Key messages:

- A significant linear correlation was established between high black tea consumption and low diabetes prevalence.
- These results are consistent with biological and physiological studies conducted on the potential effect of black tea on diabetes and obesity.
- These results should support further causality research regarding the health benefits of BT consumption on type 2 diabetes prevalence in the world.

Strengths and limitations:

- These original study results are consistent with biological and physiological studies conducted on the potential effect of black tea on diabetes and obesity. The results confirm the finding of a recent ecoogical study carried out in Europeean countries. We believe that this multidimensional approach provides valuable additional scientific information, as our findings, establishing a strong correlation between high BT consumption and low diabetes prevalence, can be considered to provide a contribution to existing biological, physiological and epidemiological studies conducted on tea consumption, diabetes and obesity.
- Diabetes prevalence data were obtained from the World Health Survey implemented by the World Health Organization, which constitutes an official source of key morbidity indicators around the world. However, the quality of data collection can be expected to be heterogeneous around the world and diabetes diagnostic criteria can vary from country to country.
- Another important concern is the interpretation of the established statistical relationship between BT consumption and diabetes prevalence. The numerous factors contributing to the growth of diabetes and obesity throughout the world confirm that "correlation does not imply causality", and that a significant linear correlation between BT consumption and diabetes prevalence does not imply that low BT consumption could *cause* diabetes. A correlation can only indicate a potential direct or indirect cause, which then needs to be further investigated.
- A frequent criticism of the use of data mining is based on the confusion between *data mining* and *data dredging* techniques. While a data mining approach is based on searching for combinations of variables that might show potential correlations, data dredging can generate misleading results. When a number of hypotheses are tested, it is expected that some will falsely appear to be statistically significant, since every database can contain potential random correlations. A robust data mining approach must therefore always be based on a clear research strategy and a limited number of relevant meaningful assumptions.

Background

Various study designs have been used to assess the potential benefits of tea. As tea is the most widely used ancient hot beverage in the world, the simple act of putting tea leaves into hot water has provided ancient societies with a tasty beverage associated with the observation of certain medicinal benefits. Two principal varieties of the species are used: the small-leaved Chinese variety (*C. sinensis sinensis*), also used for green tea and white tea, and the large-leaved Assamese variety (*C. sinensis assamica*), which has been traditionally used only for black tea (BT). Ancient Chinese civilizations realised that by using a special fermentation process, tea leaves would become darker allowing them to be stored for longer periods of time. During this fermentation process, in which green tea oxidises to form black tea, caffeine tends to remain constant, while the types of flavonoids present in the tea differ. Green tea contains simple flavonoids called catechins, whilst BT contains complex flavonoids called theaflavins and thearubigins, which could be the chemical entities responsible for a number of potential health benefits. These tea types were called black tea because of the change in colour of the leaves as a result of this fermentation process. 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 [1-9]. 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 [10], protect against oxidative damage and apoptosis in human bronchial epithelial cells induced by tobacco 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 [11] and appear to exert a protective activity against *Vibrio cholerae*[12]. However, research into the potential beneficial effects of tea appears to be most active in the field of cardiovascular diseases, in view of the number 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 [13]. Moreover, dietary consumption of tea catechins would have beneficial effects on blood pressure and lipid parameters [14]. 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 randomised 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 [15]. Anti-diabetes properties of BT have been suggested by several very specific studies, such as a change in pancreatic function in streptozotocin-induced glucose-intolerant rats [16, 17], but also in some human studies investigating other hot beverages [18-21]. The relatively recent interest in BT may be explained by the fact that BT is historically the type of tea most widely consumed in Western countries, probably due to its good storage properties, promoting active trade with tea-producing countries in Asia. Although there has recently been a renewed interest in green tea in industrialized countries, BT represents over ninety percent of all tea sold in the West. Despite the number of publications investigating the effects of tea components and green tea in particular, large-scale clinical dose-effect studies are still lacking and it is difficult to interpret the clinical significance of results derived from some biological studies. Considerably fewer studies have been conducted specifically on BT, mostly investigating its antioxidant properties [22, 23], and cardiovascular effects [24, 25]. It could therefore be interesting to investigate some key dietary habits in relation with lifestyle and health effects at a global level, in view of the perceived positive health effects of black tea (BT), which have been described for centuries [26, 27]. Because of 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 using advanced exploratory statistical methods. 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. An almost 6-fold increase in the number of people with diabetes has been observed over the last few decades. The International Diabetes Federation (IDF) reports that the number of people with diabetes will escalate from 285 million to 438 million between 2010 and 2030 [1] and the number of persons with IGT will increase from 344 to 472 million. By 2030, there will be over 900 million people worldwide with diabetes or at high risk of diabetes. Diabetes confers about a two-fold excess risk for a wide range of vascular diseases [2]. Furthermore, diabetic retinopathy is a common and specific microvascular complication of diabetes, and remains the leading cause of preventable blindness in working-aged people [3]. With one of the highest prevalences of all human diseases, diabetes is now a global epidemic with devastating health, social and economic consequences [4]. In certain ethnic groups, such as Asian populations, diabetes develops at a younger age than in Caucasian populations. Several distinctive features are apparent in the pathogenic factors for diabetes and their thresholds in Asian populations [5]. In conjunction with genetic susceptibility, type 2 diabetes is brought on by environmental and behavioural factors such as a sedentary lifestyle, overly rich nutrition and obesity and results in a huge economic burden [6]. It could therefore be interesting to investigate some key dietary habits in relation with lifestyle and health effects at a global level. For example, the positive health effects of black tea (BT) have been observed for centuries [7, 8].

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.

Formatted: Space After: 0 pt, Adjust space between Latin and Asian text, Adjust space between Asian text and numbers

~~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.~~

Material and Method

Data sources

BT consumption data were derived from a specific international trade survey compiling sales data conducted in 2009 by Euromonitor International, an independent agency specialized in market research [28]~~[9]~~. Consumption data are derived from black tea international trading registries, used by black tea importers to adapt international orders to local sales. Yearly consumption data expressed in kilograms per capita were available for the following 50 countries: Argentina, Australia, Austria, Belgium, Brazil, Bulgaria, Canada, Chile, China, Colombia, Czech Republic, Denmark, Egypt, Finland, France, Germany, Greece, Hungary, India, Indonesia, Ireland, Israel, Italy, Japan, Malaysia, Mexico, Morocco, Netherlands, New Zealand, Norway, Philippines, Poland, Portugal, Romania, Russia, Saudi Arabia, Singapore, Slovakia, South Africa, South Korea, Spain, Sweden, Switzerland, Thailand, Turkey, Ukraine, United Kingdom, USA, Venezuela, Vietnam (Figure 1). Highest BT consumptions (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), as the Chinese population drinks 30 times more green tea (0.036 kg per inhabitant) than black tea.

Field Code Changed

Epidemiological data were derived from a specific analysis of the World Health Survey (WHS) conducted by the World Health Organization (WHO). Each year, the WHS compiles comprehensive baseline information on the health of populations and health system outcomes [29]~~[10]~~. Using the 2009 dataset (table 1), ~~-~~five key health indicators were selected in 50 countries in both men and women for all age groups: prevalence of respiratory diseases, prevalence of infectious diseases (tuberculosis and HIV), prevalence of cancer, prevalence of cardiovascular diseases and prevalence of diabetes ~~(Figure 2)~~.

Field Code Changed

Methods

~~Data analyses were based on a systematic data mining approach. Data mining (sometimes called data or knowledge discovery) is generally defined as the process of analysing data from different perspectives and summarising these data into meaningful information. This approach is useful to analyse data derived from different dimensions or perspectives and to detect potential relationships between variables. Technically, data mining consists of discovering specific correlations or patterns in large relational databases. Data mining combines methods from statistics and artificial intelligence with database management and is considered to be an increasingly important tool. It is currently used in a wide range of scientific applications in~~ health [30-33]~~[11-14]~~.

Field Code Changed

In this study, the data mining approach used 3 phases: "firstly, a "calibrated principal component analyses" (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 as the mathematical projection of the 6 dimensions defined by the 6 variables into 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."

Formatted: Line spacing: 1.5 lines

~~firstly, a "calibrated principal component analyses" (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.~~

Normative principal component analysis (PCA)

PCA is a mathematical procedure that uses mathematical projections to convert a set of *n* possibly correlated variables representing *n* dimensions into a smaller number of dimensions called "principal components" classically represented in 2 or 3 axes F1, F2, F3. The projections use orthogonal transformations defined in such a way that the first principal component (first axis) has the highest possible variance in order to synthesize most of the initial information. The main objective of PCA is to reduce the dimensionality of the data set. PCA is often presented as a technique of factor analysis for quantitative variables. Multiple Correspondence Analysis (MCA) is another type of factor analysis for quantitative, qualitative and categorical variables and is useful to conduct multi-criteria analyses such as multi-criteria risk assessment [34]~~[15]~~. A "normative PCA" was selected for our study, as the 6 variables (BT consumption per capita and 5 key health indicators) are quantitative variables and this analysis was calibrated to study potential correlations.

Field Code Changed

Correlation circle

The correlation circle shows a projection of the initial variables in a dimensional space represented by axes F1 and F2 [35]~~[16]~~. Variables are presented as vectors from the centre. When two vectors are close to the correlation circle, they can be: i) close to each other, meaning a positive correlation ii) orthogonal from each other, meaning that they are not correlated iii) on the opposite side from the centre, meaning a significant negative correlation. When some vectors are closed to the centre, this means that some information is carried on other axes, and that any interpretation might be hazardous. This can be confirmed by looking at another correlation circle constructed with axes F1 and F3 or with axes F2 and F3. The correlation circle is then used to identify the potential proximity with the 6 vectors and to assess their potential correlations. Should a vector representing the variable "BT consumption" be closed to the correlation circle and point to a similar direction compared to any of the other 5 vectors representing health indicators, this would indicate a positive correlation between the two variables.

Field Code Changed

Linear correlation model

Once identified by the correlation circle, potential correlations between BT consumption and one or more health indicators can be described using key statistical parameters, such as the coefficient r^2 and the statistical significance p. Using a linear correlation model between BT consumption and one health indicator then determines the extent to which the values of these two variables are potentially "proportional" to each other (BT consumption increases or decreases with one specific disease prevalence). The linear model formula is: **y = ax + b** (y= health indicator; x = BT consumption; a and b are the model calculated coefficients). The structure of this formula suggests that the variables x and y are linearly related and thus proportional; that is, the correlation is high if it can be represented by a straight line (upwards or downwards slope). If so, this line will represent the linear model, also called a "regression line" or "least squares line" because the sum of the squared distances of all the data points from the line is the lowest possible. The coefficient r^2 (coefficient of determination) represents the proportion of common variations between the two variables and establishes the "strength" of the relationship. In order to evaluate the potential correlation between BT consumption and one specific health indicator, it is therefore important to know r^2, the statistical significance p of the correlation (calculated by a Fisher-Snedecor test) and the statistical significance of the difference to 0 of the coefficient "a" (Student's t-test).

Results

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.

The "correlation circle" (Figure 2) shows that the "BT consumption" vector was strictly opposite the "Diabetes prevalence" vector, establishing a strong statistical negative correlation. Vectors concerning the other key health indicators (infectious diseases, respiratory diseases, cancer and cardiovascular diseases) were represented with a large angle (close to orthogonal) compared to the BT vector, confirming poor statistical relationships between BT and these 4 Health indicators. Of particular interest was the interpretation of the "infectious disease" vector, which seemed to be close to the BT vector in a two dimensions projection , 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. 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.

Then linear correlation model with the format y= ax+b is represented by the following formula and is presented in Figure 3:

Diabetes prevalence = - 0.0171183 * BT consumption + 6173.64

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 be unknown (6,173 cases per 100,000 inhabitants). Based on 42 countries, the p value of the Fisher-Snedecor test was 0.003, which is highly significant, confirming the relevance of the linear model. The coefficient r^2 was equal to 0.199. Student's t-test confirmed that the slope coefficient (0.0171183) was significantly different from 0 (p=0.003) with a 5% confidence interval between [-0.028; -0.006]. The negativity of coefficient "a" indicates that diabetes prevalence decreases as BT consumption increases, confirming a negative correlation.

~~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.~~

~~The "correlation circle" (Figure 3) shows that the "BT consumption" vector was strictly opposite the "Diabetes prevalence" vector, establishing a strong statistical negative correlation. Vectors concerning the other key health indicators (infectious diseases, respiratory diseases, cancer and cardiovascular diseases)~~

~~were represented with a large angle (close to orthogonal) compared to the BT vector, confirming poor statistical relationships between BT and these 4 Health indicators. 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. 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.~~

~~The linear correlation model can be expressed as follows:~~

~~**Diabetes prevalence = a * BT consumption + b**~~

~~Based on 42 countries, the p value of the Fisher-Snedecor test was 0.01, which is highly significant, confirming the relevance of the linear model. The coefficient r^2 was equal to 0.501.~~

~~The coefficient a = -0.0171183 and a Student's t-test confirmed that this coefficient was significantly different from 0 (p=0.001) with a 5% confidence interval between [-0.007; -0.027]. The negativity of the coefficient "a" means that when BT consumption increases, diabetes prevalence decreases, confirming a negative correlation (42).~~

~~Then linear correlation model can be represented by the following formula and is presented in Figure 4::~~

~~**Diabetes prevalence = - 0.0171183 * BT consumption + 6173.64**~~

Discussion

Limitations

Formatted: Font: 11 pt, Bold, Italic
Formatted: Normal (Web), Space After: 0 pt
Formatted: Font: Bold, Italic

This study establishes an inverse linear statistical relationship between high BT consumption and diabetes prevalence in the world, and confirms the findings of the European ecological study establishing a similar relationship~~This study establishes, for the first time, a linear statistical relationship between high BT consumption and low diabetes prevalence in the countries that formed the basis for this analysis.~~ As in any database analysis, the very first limitation of this study is related to the quality of the data. WHO prevalence data were obtained from the WHS, which constitutes a convenient and official source of key morbidity indicators around the world. The general design of the WHS is based on population sampling organized in the 192 Member States of the United Nations using face-to-face or telephone interviews. As the survey questionnaire offers a menu of choices of modules for each country, and lets the country select the survey approach (Household face-to-face survey, Computer-Assisted Telephone Interview or

Formatted: Font: 11 pt

Computer-Assisted Personal Interview), the quality of data collection can be expected to be heterogeneous around the world.

Furthermore, 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. On the other hand, any fixed survey design with fixed criteria would not be appropriate everywhere, for example in countries with low telephone network coverage when planning telephone interviews. Other approaches to estimate prevalence of diabetes in the world have been studies using literature and data extrapolations [36], confirming the growing burden of diabetes.

Another important concern is the interpretation of the established statistical relationship between BT consumption and diabetes prevalence. Using advanced data mining techniques, we tested the potential statistical relationship between BT consumption and 5 health indicators, without any *a priori* assumptions in relation to any of these health indicators. We observed that, among the 5 health indicators, only the "prevalence of diabetes" indicator appeared to have a strong statistical relationship with BT consumption. 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. A classical 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 the basis for any individual assumptions. However, when interesting and 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 studies was that published by Keys in 1980 [37] concerning the relationship with dietary habits and coronary heart disease in 7 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 [38]. For these reasons, ecological studies can be very useful for international comparisons, while case-control studies are exclusively based on local information. Furthermore, when

Formatted: Font color: Text 1

Field Code Changed

Formatted: Superscript

Formatted: Superscript

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.

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. ***Growing interest of food components that may support weight management and glucose metabolism***

Formatted: Font: Bold, Italic

Formatted: Font: Bold, Italic

Our results confirm the recent publication from the InterAct Consortium which has carried out a European ecological study and has confirmed a linear inverse association between tea consumption and incidence of type 2 diabetes in Europe [39]. Various study designs have already been used to assess the potential benefits of tea. As this is the most widely used ancient hot beverage in the world, the simple act of putting tea leaves into hot water has provided ancient societies with a tasty beverage associated with the observation of certain medicinal benefits. Two principal varieties of the species are used: the small-leaved Chinese variety (*C. sinensis sinensis*), also used for green tea and white tea, and the large-leaved Assamese variety (*C. sinensis assamica*), which has been traditionally used only for BT. Ancient Chinese civilizations realised that using a special fermentation process, tea leaves would become darker allowing them to be stored for longer periods of time. During this fermentation process, in which green tea oxidises to form black tea, caffeine tends to remain constant, while the types of flavonoids present in the tea differ. Green tea contains simple flavonoids called catechins, whilst BT contains complex flavonoids called theaflavins and thearubigins, which could be the chemical entities responsible for a number of potential health benefits. These tea types were called black tea because of the change in colour of the leaves as a result of this fermentation process. Numerous *in vitro* and *in vivo* studies have demonstrated the health benefits of green tea, mainly in cancer, cardiovascular disease, chronic inflammation or cognitive functions [17-25]. However, large-scale clinical dose-effect studies are still missing and it is difficult to interpret the clinical significance of results derived from some biological studies. Considerably fewer studies have been conducted on BT, mostly investigating its antioxidant properties [26, 27], and cardiovascular effects [28, 29]. Anti-diabetes properties of BT have been suggested by several very specific studies, such as a change in pancreatic function in streptozotocin-induced glucose-intolerant rats [30, 31], but also in some human studies also investigating other hot beverages [32-35]. The relatively recent interest in BT may be explained by the fact that BT is historically the type of tea most widely consumed in Western countries, probably due to its good storage properties, promoting active trade with tea-producing countries in Asia. Although there has recently been a renewed interest in green tea in industrialized countries due to its popular health benefits, BT represents over ninety percent of all tea sold in the West.

Formatted: Font: 11 pt

Formatted: Superscript

The type 2 diabetes epidemic in many countries has stimulated interest in food components that may support weight management. An almost 6-fold increase in the number of people with diabetes has been

Formatted: Don't adjust space between Latin and Asian text, Don't adjust space between Asian text and numbers observed over the last few decades. The International Diabetes Federation (IDF) reports that the number of people with diabetes will escalate from 285 million to 438 million between 2010 and 2030[36] and the number of persons with IGT will increase from 344 to 472 million. By 2030, there will be over 900 million people worldwide with diabetes or at high risk of diabetes. Diabetes confers about a two-fold excess risk for a wide range of vascular diseases[40]. Furthermore, diabetic retinopathy is a common and specific microvascular complication of diabetes, and remains the leading cause of preventable blindness in working-aged people[41]. With one of the highest prevalences of all human diseases, diabetes is now a global epidemic with devastating health, social and economic consequences[42]. In certain ethnic groups, such as Asian populations, diabetes develops at a younger age than in Caucasian populations. Several distinctive features are apparent in the pathogenic factors for diabetes and their thresholds in Asian populations[43]. In conjunction with genetic susceptibility, type 2 diabetes is brought on by environmental and behavioural factors such as a sedentary lifestyle, overly rich nutrition and obesity and results in a huge economic burden[44]. 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[45~~36~~].

Although many laboratory studies have observed physiological effects of BT on glucose metabolism[17, 18, 46, 47~~31, 32, 37, 38~~], the underlying mechanisms remain unclear. The results of human intervention studies are mixed[48~~39~~] and the role of caffeine has been suggested but not clearly established[47, 49~~38, 40~~]. Neyestani *et al*[46~~37~~] found that regular daily intake of BT improves oxidative stress biomarkers and decreases serum C-reactive protein levels in type 2 diabetic patients. Histological studies on pancreas cells published by Manikandan *et al*[17~~31~~] concluded that the BT extract contributes to regeneration of damaged pancreas cells and protects pancreatic beta cells by its antioxidant action. Nonetheless, the role of environment, dietary and lifestyle practices is fundamental when comparing health indicators around the world. Psaltopoulou *et al*[50~~41~~] confirmed that low-glycaemic index dietary patterns reduced both fasting blood glucose and glycated proteins independently of carbohydrate consumption. Diets rich in whole-grain, cereal high-fibre products, and non-oil-seed pulses would also be beneficial. As vitamins and minerals play an important role in glucose metabolism, understanding the impact of potential vitamin and mineral deficiencies across cultures is also relevant to better organization of prevention and management of type 2 diabetes[51, 52~~42, 43~~]. An observational study based on nearly 37,000 middle-aged Chinese reported a 14% reduction in the risk of developing type 2 diabetes by drinking one or more cups of tea per day[53~~44~~]. This was confirmed by two meta-analyses published by Huxley *et al*[47~~38~~] and Jing *et al*[54~~45~~]. Flavonoids are believed to support normal glucose metabolism via anti-inflammatory effects and increased insulin activity[55, 56~~46, 47~~]. 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[57~~48~~]. An observational study of 4,300 Dutch adults found that flavonoid intake was highest in women who gained the least weight over a 14-year period[58~~49~~]. Furthermore, as physical activity with or without diet

Formatted: Superscript

Formatted: Superscript

Formatted: Superscript

Formatted: Superscript

Formatted: Superscript

Formatted: Superscript

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

contributes to a healthier lifestyle, this important factor must be considered when comparing health indicators between industrialized and emerging countries. Given rapid population growth, increased urbanization, and adverse lifestyle changes, the obesity/type 2 diabetes epidemic in resource-poor nations was predicted in the 1990s and has now been fully confirmed [59~~50~~], underlying the importance of a better understanding of predictive and potentially protective factors.

Field Code Changed

Correlation and causality

Formatted: Font: Bold, Italic

The number of factors contributing to the growth of diabetes and obesity in the world confirms that "correlation does not imply causality ", and that a significant linear correlation between BT consumption and diabetes prevalence does not imply that low BT consumption could *cause* diabetes. If one factor is established as causing another, then the two factors are most certainly correlated. However, the opposite cannot be concluded. Thus, a correlation can only indicate a potential direct or indirect possible cause, which then needs to be further investigated. This paradigm and the connotations of causality may be the most important considerations affecting biostatistics, not only in ecological studies but also in major epidemiological study designs [60]. A well known example of epidemiological cause-and-effect misinterpretations is the correlation that was established between hormone replacement therapy and a lower incidence of coronary heart disease. This association has been more recently explained by the fact that women taking hormone replacement therapy were more likely to come from higher socio-economic levels, which could explain the lower incidence of coronary heart disease [61]. Establishing causality is one of the most difficult challenges in public health. For instance, in clinical research, randomised controlled clinical trials are performed to establish potential significant differences between two groups. However, establishing a difference is not a demonstration of causality. Another example is case-control studies, which compare individuals with a specific disease ("cases") with a group of individuals without the disease ("controls"). An association between the hypothesized exposure and the disease studied would be reflected by a higher proportion in exposed cases, but this cannot constitute a real demonstration of causality. A potential causality can only be established with the convergence of interdisciplinary scientific evidence (biological, physiological, epidemiological, etc.) and reasonable explanations based on longitudinal studies. In any case, ecological research can address important issues that cannot be easily addressed by other study designs. Ecological studies are frequently used when alternative study designs are not possible (eg, randomised control trials), such as when investigating the effect of geographical factors on disease incidence. Our research, like all ecological studies and most other epidemiological approaches,presents a number of limitations 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

Formatted: Superscript

Formatted: Superscript

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. Despite the fact that a number of biological, physiological and epidemiological field studies have provided evidence linking BT consumption and glucose metabolism [16, 17, 22, 46-48, 55, 56], a large-scale randomised controlled trial of tea consumption and diabetes risk would be useful to confirm these findings.

Formatted: Superscript

The number of factors contributing to the growth of diabetes and obesity in the world confirm that "correlation does not imply causality", and that a significant linear correlation between BT consumption and diabetes prevalence does not imply that low BT consumption could *cause* diabetes. If one factor is established as causing another, then the two factors are most certainly correlated. However, the opposite cannot be concluded. Thus, a correlation can only indicate a potential direct or indirect possible cause, which then needs to be further investigated. This paradigm and the connotations of causality may be the most important considerations affecting biostatistics in major epidemiological study designs [51]. A well known example of epidemiological cause-and-effect misinterpretations is the correlation that was established between hormone replacement therapy and a lower incidence of coronary heart disease. This association has been more recently explained by the fact that women taking hormone replacement therapy were more likely to come from higher socio-economic levels, which could explain the lower incidence of coronary heart disease [52]. Establishing causality is one of the most difficult challenges in public health. For instance, in clinical research, randomized controlled clinical trials are performed to establish potential significant differences between two groups. However, establishing a difference is not a demonstration of causality. Another example is case-control studies, which compare individuals with a specific disease ("cases") with a group of individuals without the disease ("controls"). An association between the hypothesized exposure and the disease studied would be reflected by a higher proportion in exposed cases, but this cannot constitute a real demonstration of causality. A potential causality can only be established with the convergence of interdisciplinary scientific evidence (biological, physiological, epidemiological, etc.) and reasonable explanations based on longitudinal studies.

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.~~[26, 30, 31, 37-39, 46, 47] ***Data mining and data dredging***

~~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.~~

Beyond the causality issue, a frequent criticism of using data mining was based on the confusion between *data mining* and *data dredging* techniques. While a data mining approach is based on searching for combinations of variables that might show potential correlations, data-dredging (also called "data fishing") can generate misleading results[62]. When a number of hypotheses are tested, it is expected that some will falsely appear to be statistically significant, since every database can contain potential random correlations. A robust data mining approach must therefore always be based on a clear research strategy and a limited number of relevant meaningful assumptions. In our assessment, we used a systematic data mining approach to test potential correlations between 6 selected variables (BT consumption and 5 key health indicators). PCA was used to describe and structure the dataset before testing any correlations. In our study, only one linear correlation model was constructed between BT consumption and diabetes prevalence, based on the most relevant association suggested by the PCA. This consistent approach is quite different from screening numerous cross-regression analyses between all variables of one particular dataset. The data mining approach can be considered to be a "radar tracking system", allowing detection, tracking and classification of potential "targets" in the framework of a particular environment. This is particularly useful when exploring complex databases, as data mining can identify original statistical evidence, which would never be discovered by means of classical statistical techniques. As an example, the significant progress in genomics would not have been possible without the use of data mining techniques. Despite the data collection homogeneity issue inherent to large cross-country comparisons, we believe that this multidimensional approach can provide valuable additional scientific information, completing published biological, physiological and epidemiological studies conducted on tea consumption, diabetes and obesity. These results should support further causality research regarding the health benefits of BT consumption on type 2 diabetes prevalence in the world.

Formatted: Font: Times New Roman

Formatted: French (France)

Formatted: Font: Bold, Italic

Formatted: Normal (Web), Space After: 0 pt

Formatted: Superscript

Beyond the causality issue, a frequent criticism of using data mining was based on the confusion between *data mining* and *data dredging* techniques. While a data mining approach is based on searching for combinations of variables that might show potential correlations, data-dredging (also called "data fishing") can generate misleading results.[53]. When a number of hypotheses are tested, it is expected that some will falsely appear to be statistically significant, since every database can contain potential random correlations. A robust data mining approach must therefore always be based on a clear research strategy and a limited number of relevant meaningful assumptions. In our assessment, we used a systematic data mining approach to test potential correlations between 6 selected variables (BT consumption and 5 key health indicators). PCA was used to describe and structure the dataset before testing any correlations. In our study, only one linear correlation model was constructed between BT consumption and diabetes prevalence, based on the most relevant association suggested by the PCA. This consistent approach is quite different from screening numerous cross-regression analyses between all variables of one particular dataset. The data mining approach can be considered to be a "radar tracking system", allowing detection, tracking and classification of potential "targets" in the framework of a particular environment. This is particularly useful when exploring complex databases, as data mining can identify original statistical evidence, which would never be discovered by means of classical statistical techniques. As an example, the significant progress in genomics would not have been possible without the use of data mining techniques.[54]. research regarding the health benefits of BT consumption on type 2 diabetes prevalence in the world.

Despite the data collection homogeneity issue inherent to large cross-country comparisons, we believe that this multidimensional approach could provide valuable additional scientific information, which is why our findings establishing a strong correlation between high BT consumption and low diabetes prevalence in these countries should be considered as a contribution to existing biological, physiological and epidemiological studies conducted on tea consumption, diabetes and obesity. These results should support further causality research regarding the health benefits of BT consumption on type 2 diabetes prevalence in the world.

Formatted: Font: Times New Roman

Formatted: Font: Times New Roman

Formatted: Font: Times New Roman

Formatted: Font: Times New Roman

Formatted: Font: Times New Roman

Formatted: Font: Times New Roman

Formatted: Font: Times New Roman, Bold

Formatted: Font: Times New Roman

Formatted: Font: Times New Roman, Italic

Formatted: Font: Times New Roman

Formatted: Font: Times New Roman, Bold

Formatted: Font: Times New Roman

Formatted: Font: Times New Roman, Italic

Formatted: Font: Times New Roman

Formatted: Font: Times New Roman, Bold

Formatted: Font: Times New Roman

Formatted: Font: Times New Roman, Italic

Formatted: Font: Times New Roman

Formatted: Font: Times New Roman, Bold

Formatted: Font: Times New Roman

Formatted: Font: Times New Roman, Italic

Formatted: Font: Times New Roman

Formatted: Font: Times New Roman, Bold

Formatted: Font: Times New Roman

Formatted: Font: Times New Roman, Italic

Formatted: Font: Times New Roman

References

1. Kang H, Rha S, Oh K, Nam C. Green tea consumption and stomach cancer risk: a meta-analysis. *Epidemiol Health.* 2010;32:e2010001.
2. Iwasaki M, Inoue M, Sasazuki S, et al. Green tea drinking and subsequent risk of breast cancer in a population to based cohort of Japanese women. *Breast Cancer Res.* 2010;12(5):R88.
3. Lee A, Liang W, Hirayama F, Binns C. Association between green tea consumption and lung cancer risk. *J Prev Med Public Health.* 2010;43(4):366-367.
4. Moore RJ JK, Minihane AM. Green tea (Camellia sinensis) catechins and vascular function. *Br J Nutr.* 2009;102(12):1790-1802.
5. Feng L, Gwee X, Kua E, Ng T. Cognitive function and tea consumption in community dwelling older Chinese in Singapore. *J Nutr Health Aging.* 2010;14(6):433-438.

6. de Mejia E, Ramirez-Mares M, Puangpraphant S. Bioactive components of tea: cancer, inflammation and behavior. *Brain Behav Immun.* 2009;23(6):721-731.
7. Béliveau R, Gingras D. Green tea: prevention and treatment of cancer by nutraceuticals. *Lancet.* 2004;364(9439):1021-1022.
8. Walsh G. Tea and heart disease. *Lancet.* 1997;349(9053):735.
9. Ras R, Zock P, Draijer R. Tea Consumption Enhances Endothelial-Dependent Vasodilation; a Meta-Analysis. *PLoS ONE* 2011;6(3):e16974.
10. 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):435-444.
11. 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):123-132.
12. Toda M, Okubo S, Ikigai H, et al. The protective activity of tea catechins against experimental infection by Vibrio cholerae O1. *Microbiol Immunol.* 1992;36(9):999-1001.
13. Islam MA. Cardiovascular effects of green tea catechins: progress and promise. . *Recent Pat Cardiovasc Drug Discov.* 2012;7(2):88-99.
14. Hodgson JM, Puddey IB, Woodman RJ, et al. Effects of black tea on blood pressure: a randomized controlled trial. *Arch Intern Med.* 2012;172(2):186-188.
15. 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):1259-1270.
16. Dias T, Bronze MR, Houghton PJ, Mota-Filipe H, Paulo A. The flavonoid-rich fraction of Coreopsis tinctoria promotes glucose tolerance regain through pancreatic function recovery in streptozotocin-induced glucose-intolerant rats. *J Ethnopharmacol.* 2010;132(2):483-490.
17. Manikandan R SR, Thiagarajan R, Sivakumar MR, Meiyalagan V, Arumugam M. Effect of black tea on histological and immunohistochemical changes in pancreatic tissues of normal and streptozotocin-induced diabetic mice (Mus musculus). *Microsc Res Tech.* 2009;72(10):723-726.
18. Oba S NC, Nakamura K, Fujii K, Kawachi T, Takatsuka N, Shimizu H. Consumption of coffee, green tea, oolong tea, black tea, chocolate snacks and the caffeine content in relation to risk of diabetes in Japanese men and women. *Br J Nutr.* 2010;103(3):453-459.
19. Isogawa A, Noda M, Takahashi Y, Kadowaki T, Tsugane S. Coffee consumption and risk of type 2 diabetes mellitus. *Lancet.* 2003;361(9358):703-704.
20. Yoshioka K, Kogure A, Yoshida T, Yoshikawa T. Coffee consumption and risk of type 2 diabetes mellitus. *Lancet.* 2003;361(9358):703.
21. Reunanen A, Heliövaara M, Aho K. Coffee consumption and risk of type 2 diabetes mellitus. *Lancet.* 2003;361(9358):702-703.
22. Pękal A, Dróżdż P, Biesaga M, Pyrzynska K. Evaluation of the antioxidant properties of fruit and flavoured black teas. *Eur J Nutr.* 2011;Mar 1.
23. Adhikary B, Yadav S, Roy K, Bandyopadhyay S, Chattopadhyay S. Black tea and theaflavins assist healing of indomethacin-induced gastric ulceration in mice by antioxidative action. *Evid Based Complement Alternat Med.* 2011(Sep 29): pii: 546560.
24. Bahorun T, Luximon-Ramma A, Gunness TK, et al. Black tea reduces uric acid and C-reactive protein levels in humans susceptible to cardiovascular diseases. *Toxicology.* 2010;278(1):68-74.
25. Tokudome S, Nahomi I, Goto C, Tokudome Y, Moore M. Black tea and cardiovascular disease. *Int J Epidemiol.* 2005;34(2):482-483.
26. 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.
27. 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.
28. Euromonitor. Hot Drinks: trade sources. 2010;www.euromonitor.com.
29. WHO. Global Health Survey. 2009;http://www.who.int/healthinfo/survey/en/.
30. Naqa I, Deasy J, Mu Y, et al. Datamining approaches for modeling tumor control probability. *Acta Oncol.* 2010;49(8):1363-1373.

Formatted ... [1]
Formatted ... [2]
Formatted ... [3]
Formatted ... [4]
Formatted ... [5]
Formatted ... [6]
Formatted ... [7]
Formatted ... [8]
Formatted ... [9]
Formatted ... [10]
Formatted ... [11]
Formatted ... [12]
Formatted ... [13]
Formatted ... [14]
Formatted ... [15]
Formatted ... [16]
Formatted ... [17]
Formatted ... [18]
Formatted ... [19]
Formatted ... [20]
Formatted ... [21]
Formatted ... [22]
Formatted ... [23]
Formatted ... [24]
Formatted ... [25]

31. Zhang F, Chen J. Data mining methods in omics-based biomarker discovery. *Methods Mol Biol.* 2011(719):511-526.
32. Wei CK SS, Yang MC. Application of Data Mining on the Development of a Disease Distribution Map of Screened Community Residents of Taipei County in Taiwan. *J Med Syst.* 2011(Feb 25).
33. Harpaz R, Haerian K, Chase H, Friedman C. Statistical Mining of Potential Drug Interaction Adverse Effects in FDA's Spontaneous Reporting System. *AMIA Annu Symp Proc.* 2010(Nov 13):281-285.
34. Briand S, Beresniak A, Nguyen T, et al. Assessment of yellow fever epidemic risk: an original multi-criteria modeling approach. *PLoS Negl Trop Dis.* 2009;3(7):e483.
35. Everitt B, Dunn G. Applied Multivariate data analysis. *Lavoisier Publisher.* 2001, 2d ed.:320p
36. Shaw JE, Sicree RA, Zimmet PZ. Global estimates of the prevalence of diabetes for 2010 and 2030. *Diabetes Research and Clinical Practice.* 2010(87):4-14.
37. Keys A. Seven Countries: A Multivariate Analysis of Death and Coronary Heart Disease. *Harvard University Press.* 1980.
38. 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;331(7510):189-191.
39. Consortium. TI. Tea Consumption and Incidence of Type 2 Diabetes in Europe: The EPIC-InterAct Case-Cohort Study. *PLoS One.* 2012(7):5.
40. Emerging Risk Factors Collaboration, Sarwar N, Gao P, et al. Diabetes mellitus, fasting blood glucose concentration, and risk of vascular disease: a collaborative meta-analysis of 102 prospective studies. *Lancet.* 2010;375(9733):2215-2222.
41. Cheung N, Mitchell P, Wong TY. Diabetic retinopathy. *Lancet.* 2010;376(9735):124-136.
42. Danaei G, Finucane MM, Lu Y, et al. National, regional, and global trends in fasting plasma glucose and diabetes prevalence since 1980: systematic analysis of health examination surveys and epidemiological studies with 370 country-years and 2•7 million participants. *Lancet.* 2011;10.1016/S0140-6736(11):60679-X.
43. Ramachandran A MR, Snehalatha C. Diabetes in Asia. *Lancet.* 2010;375(9712):408-418.
44. Zimmet P, Alberti KG, Shaw J. Global and societal implications of the diabetes epidemic. *Nature Reviews.* 2001;414(6865):782-787.
45. Ang YG, Wu XC, Toh MP, Chia KS, Heng BH. Progression Rate of newly diagnosed Impaired Fasting Glycemia to Type 2 Diabetes Mellitus: a study using the National Healthcare Group Diabetes Registry in Singapore. *J Diabetes.* 2011;Nov 7.
46. Neyestani T, Shariatzade N, Kalayi A, et al. Regular daily intake of black tea improves oxidative stress biomarkers and decreases serum C-reactive protein levels in type 2 diabetic patients. *Ann Nutr Metab.* 2010;57(1):40-49.
47. Huxley R, Lee C, Barzi F, et al. Coffee, decaffeinated coffee, and tea consumption in relation to incident type 2 diabetes mellitus: a systematic review with meta-analysis. *Arch Intern Med.* 2009;169(22):2053-2063.
48. Hayashino Y, Fukuhara S, Okamura T, Tanaka T, Ueshima H, Group. H-OR. High oolong tea consumption predicts future risk of diabetes among Japanese male workers: a prospective cohort study. *Diabet Med.* 2011(Jan 18).
49. Goto A, Song Y, Chen B, Manson J, Buring J, Liu S. Coffee and caffeine consumption in relation to sex hormone-binding globulin and risk of type 2 diabetes in postmenopausal women. *Diabetes.* 2011;60(1):269-275.
50. Psaltopouloum T, Ilias I, Alevizaki M. The role of diet and lifestyle in primary, secondary, and tertiary diabetes prevention: a review of meta-analyses. *Rev Diabet Stud. .* 2010;7(1):26-35.
51. Martini L, Catania A, Ferreira S. Role of vitamins and minerals in prevention and management of type 2 diabetes mellitus. *Nutr Rev.* 2010;68(6):341-354.
52. Suksomboon N, Poolsup N, Sinprasert S. Effects of vitamin E supplementation on glycaemic control in type 2 diabetes: systematic review of randomized controlled trials. *J Clin Pharm Ther.* 2011;36(1):53-63.

Formatted ... [26]
Formatted ... [27]
Formatted ... [28]
Formatted ... [29]
Formatted ... [30]
Formatted: Indent: Left: 0", First line: 0"
Formatted ... [31]
Formatted ... [32]
Formatted ... [33]
Formatted ... [34]
Formatted ... [35]
Formatted ... [36]
Formatted ... [37]
Formatted ... [38]
Formatted ... [39]
Formatted ... [40]
Formatted ... [41]
Formatted ... [42]
Formatted ... [43]
Formatted ... [44]
Formatted ... [45]
Formatted ... [46]
Formatted ... [47]

53. Odegaard A, Pereira M, Koh W, Arakawa K, Lee H, Yu M. Coffee, tea and incident type 2 diabetes: the Singapore Chinese Health Study. *American Journal of Clinical Nutrition.* 2008;88(4):979-985.
54. Jing Y, Han G, Hu Y, Bi Y, Li L, Zhu D. Tea consumption and risk of type 2 diabetes: a metaanalysis of cohort studies. *J Gen Intern Med.* . 2009;24(5):557-562.
55. Nicolle E, Souard F, Faure P, Boumendjel A. Flavonoids as promising lead compounds in type 2 diabetes mellitus: molecules of interest and structure-activity relationship. *Curr Med Chem.* 2011;18(17):2661-2672.
56. Miyata Y, Tanaka H, Shimada A, et al. Regulation of adipocytokine secretion and adipocyte hypertrophy by polymethoxyflavonoids, nobiletin and tangeretin. *Life Sci.* . 2011;88(13-14):613-618.
57. MS. W-P. Green tea catechins, caffeine and body-weight regulation. *Physiol Behav.* 2010;100(1):42-46.
58. Hughes L, Arts I, Ambergen T, et al. Higher dietary flavone, flavonol, and catechin intakes are associated with less of an increase in BMI over time in women: a longitudinal analysis from the Netherlands Cohort Study. *Am J Clin Nutr.* 2008;88(5):1341-1352.
59. Nour N. Obesity in resource-poor nations. *Rev Obstet Gynecol.* . 2010;3(4):180-184.
60. Ortega Calvo M, Román Torres P, Lapetra Peralta J. Epistemology as health research propedeutics. *Gac Sanit.* 2011.
61. Lawlor D, Davey Smith G, Ebrahim S. Commentary: the hormone replacement-coronary heart disease conundrum: is this the death of observational epidemiology? *Int J Epidemiol.* 2004;33(3):464-467.
62. Lord S, Gebski V, Keech A. Multiple analyses in clinical trials: sound science or data dredging? *Med J Aust.* . 2004;18(18).

1. Shaw JE, Sicree RA, Zimmet PZ. Global estimates of the prevalence of diabetes for 2010 and 2030. *Diabetes Research and Clinical Pratice.* 2010(87):4-14.
2. Emerging Risk Factors Collaboration, Sarwar N, Gao P, et al. Diabetes mellitus, fasting blood glucose concentration, and risk of vascular disease: a collaborative meta-analysis of 102 prospective studies. *Lancet.* 2010;375(9733):2215-2222.
3. Cheung N, Mitchell P, Wong TY. Diabetic retinopathy. *Lancet.* 2010;376(9735):124-136.
4. Danaei G, Finucane MM, Lu Y, et al. National, regional, and global trends in fasting plasma glucose and diabetes prevalence since 1980: systematic analysis of health examination surveys and epidemiological studies with 370 country-years and 2•7 million participants. *Lancet.* 2011;10.1016/S0140-6736(11):60679-X.
5. Ramachandran A MR, Snehalatha C. Diabetes in Asia. *Lancet.* 2010;375(9712):408-418.
6. Zimmet P, Alberti KG, Shaw J. Global and societal implications of the diabetes epidemic. *Nature Reviews.* 2001;414(6865):782-787.
7. 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.
8. 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.
9. Euromonitor. Hot Drinks: trade sources. 2010;www.euromonitor.com.
10. WHO. Global Health Survey. 2009;http://www.who.int/healthinfo/survey/en/.
11. Naqa I, Deasy J, Mu Y, et al. Datamining approaches for modeling tumor control probability. *Acta Oncol.* 2010;49(8):1363-1373.
12. Zhang F, Chen J. Data mining methods in omics-based biomarker discovery. *Methods Mol Biol.* 2011(719):511-526.
13. Wei CK SS, Yang MC. Application of Data Mining on the Development of a Disease Distribution Map of Screened Community Residents of Taipei County in Taiwan. *J Med Syst.* 2011(Feb 25).
14. Harpaz R, Haerian K, Chase H, Friedman C. Statistical Mining of Potential Drug Interaction Adverse Effects in FDA's Spontaneous Reporting System. *AMIA Annu Symp Proc.* 2010(Nov 13):281-285.

Formatted: Font: Times New Roman, Bold
Formatted: Font: Times New Roman
Formatted: Font: Times New Roman, Italic
Formatted: Font: Times New Roman
Formatted: Font: Times New Roman, Bold
Formatted: Font: Times New Roman
Formatted: Font: Times New Roman, Italic
Formatted: Font: Times New Roman
Formatted: Font: Times New Roman, Bold
Formatted: Font: Times New Roman
Formatted: Font: Times New Roman, Italic
Formatted: Font: Times New Roman
Formatted: Font: Times New Roman, Bold
Formatted: Font: Times New Roman
Formatted: Font: Times New Roman, Italic
Formatted: Font: Times New Roman
Formatted: Font: Times New Roman, Bold
Formatted: Font: Times New Roman
Formatted: Font: Times New Roman, Italic
Formatted: Font: Times New Roman
Formatted: Font: Times New Roman, Bold
Formatted: Font: Times New Roman
Formatted: Font: Times New Roman, Italic
Formatted: Font: Times New Roman
Formatted: Font: Times New Roman, Bold
Formatted: Font: Times New Roman
Formatted: Font: Times New Roman, Italic
Formatted: Font: Times New Roman
Formatted: Font: Times New Roman, Bold
Formatted: Font: Times New Roman
Formatted: Font: Times New Roman, Italic
Formatted: Font: Times New Roman
Formatted: Font: Times New Roman, Bold
Formatted: Font: Times New Roman
Formatted: Font: Times New Roman, Italic
Formatted: Font: Times New Roman
Formatted: Font: Times New Roman, Bold
Formatted: Font: Times New Roman
Formatted: Font: Times New Roman, Italic
Formatted: Font: Times New Roman
Formatted: Indent: Left: 0", First line: 0"

15. Briand S, Beresniak A, Nguyen T, et al. Assessment of yellow fever epidemic risk: an original multi-criteria modeling approach. *PLoS Negl Trop Dis.* 2009;3(7):e483.
16. Everitt B, Dunn G. Applied Multivariate data analysis. *Lavoisier Publisher.* 2001, 2d ed.:320p

Formatted: Indent: Left: 0", Hanging: 0.5"

17. Kang H, Rha S, Oh K, Nam C. Green tea consumption and stomach cancer risk: a meta-analysis. *Epidemiol Health.* 2010;32:e2010001.
18. Iwasaki M, Inoue M, Sasazuki S, et al. Green tea drinking and subsequent risk of breast cancer in a population to based cohort of Japanese women. *Breast Cancer Res.* 2010;12(5):R88.
19. Lee A, Liang W, Hirayama F, Binns C. Association between green tea consumption and lung cancer risk. *J Prev Med Public Health.* 2010;43(4):366-367.
20. Moore RJ JK, Minihane AM. Green tea (Camellia sinensis) catechins and vascular function. *Br J Nutr.* 2009;102(12):1790-1802.
21. Feng L, Gwee X, Kua E, Ng T. Cognitive function and tea consumption in community dwelling older Chinese in Singapore. *J Nutr Health Aging.* 2010;14(6):433-438.
22. de Mejia E, Ramirez-Mares M, Puangpraphant S. Bioactive components of tea: cancer, inflammation and behavior. *Brain Behav Immun.* 2009;23(6):721-731.
23. Béliveau R, Gingras D. Green tea: prevention and treatment of cancer by nutraceuticals. *Lancet.* 2004;364(9439):1021-1022.
24. Walsh G. Tea and heart disease. *Lancet.* 1997;349(9053):735.
25. Ras R, Zock P, Draijer R. Tea Consumption Enhances Endothelial-Dependent Vasodilation; a Meta-Analysis. *PLoS ONE* 2011;6(3):e16974.
26. Pękal A, Dróżdż P, Biesaga M, Pyrzynska K. Evaluation of the antioxidant properties of fruit and flavoured black teas. *Eur J Nutr.* 2011;Mar 1.
27. Adhikary B, Yadav S, Roy K, Bandyopadhyay S, Chattopadhyay S. Black tea and theaflavins assist healing of indomethacin-induced gastric ulceration in mice by antioxidative action. *Evid Based Complement Alternat Med.* 2011(Sep 29): pii: 546560.
28. Bahorun T, Luximon-Ramma A, Gunness TK, et al. Black tea reduces uric acid and C-reactive protein levels in humans susceptible to cardiovascular diseases. *Toxicology.* 2010;278(1):68-74.
29. Tokudome S, Nahomi I, Goto C, Tokudome Y, Moore M. Black tea and cardiovascular disease. *Int J Epidemiol.* 2005;34(2):482-483.
30. Dias T, Bronze MR, Houghton PJ, Mota-Filipe H, Paulo A. The flavonoid-rich fraction of Coreopsis tinctoria promotes glucose tolerance regain through pancreatic function recovery in streptozotocin-induced glucose-intolerant rats. *J Ethnopharmacol.* 2010;132(2):483-490.
31. Manikandan R SR, Thiagarajan R, Sivakumar MR, Meiyalagan V, Arumugam M. Effect of black tea on histological and immunohistochemical changes in pancreatic tissues of normal and streptozotocin-induced diabetic mice (Mus musculus). *Microsc Res Tech.* 2009;72(10):723-726.
32. Oba S NC, Nakamura K, Fujii K, Kawachi T, Takatsuka N, Shimizu H. Consumption of coffee, green tea, oolong tea, black tea, chocolate snacks and the caffeine content in relation to risk of diabetes in Japanese men and women. *Br J Nutr.* 2010;103(3):453-459.
33. Isogawa A, Noda M, Takahashi Y, Kadowaki T, Tsugane S. Coffee consumption and risk of type 2 diabetes mellitus. *Lancet.* 2003;361(9358):703-704.
34. Yoshioka K, Kogure A, Yoshida T, Yoshikawa T. Coffee consumption and risk of type 2 diabetes mellitus. *Lancet.* 2003;361(9358):703.
35. Reunanen A, Heliövaara M, Aho K. Coffee consumption and risk of type 2 diabetes mellitus. *Lancet.* 2003;361(9358):702-703.
36. Ang YG, Wu XC, Toh MP, Chia KS, Heng BH. Progression Rate of newly diagnosed Impaired Fasting Glycemia to Type 2 Diabetes Mellitus: a study using the National Healthcare Group Diabetes Registry in Singapore. *J Diabetes.* 2011;Nov 7.
37. Neyestani T, Shariatzade N, Kalayi A, et al. Regular daily intake of black tea improves oxidative stress biomarkers and decreases serum C-reactive protein levels in type 2 diabetic patients. *Ann Nutr Metab.* 2010;57(1):40-49.
38. Huxley R, Lee C, Barzi F, et al. Coffee, decaffeinated coffee, and tea consumption in relation to incident type 2 diabetes mellitus: a systematic review with meta-analysis. *Arch Intern Med.* 2009;169(22):2053-2063.

39. Hayashino Y, Fukuhara S, Okamura T, Tanaka T, Ueshima H, Group. H-OR. High oolong tea consumption predicts future risk of diabetes among Japanese male workers: a prospective cohort study. *Diabet Med.* 2011(Jan 18).
40. Goto A, Song Y, Chen B, Manson J, Buring J, Liu S. Coffee and caffeine consumption in relation to sex hormone-binding globulin and risk of type 2 diabetes in postmenopausal women. *Diabetes.* 2011;60(1):269-275.
41. Psaltopouloum T, Ilias I, Alevizaki M. The role of diet and lifestyle in primary, secondary, and tertiary diabetes prevention: a review of meta-analyses. *Rev Diabet Stud.* . 2010;7(1):26-35.
42. Martini L, Catania A, Ferreira S. Role of vitamins and minerals in prevention and management of type 2 diabetes mellitus. *Nutr Rev.* 2010;68(6):341-354.
43. Suksomboon N, Poolsup N, Sinprasert S. Effects of vitamin E supplementation on glycaemic control in type 2 diabetes: systematic review of randomized controlled trials. *J Clin Pharm Ther.* 2011;36(1):53-63.
44. Odegaard A, Pereira M, Koh W, Arakawa K, Lee H, Yu M. Coffee, tea and incident type 2 diabetes: the Singapore Chinese Health Study. *American Journal of Clinical Nutrition.* 2008;88(4):979-985.
45. Jing Y, Han G, Hu Y, Bi Y, Li L, Zhu D. Tea consumption and risk of type 2 diabetes: a metaanalysis of cohort studies. *J Gen Intern Med.* . 2009;24(5):557-562.
46. Nicolle E, Souard F, Faure P, Boumendjel A. Flavonoids as promising lead compounds in type 2 diabetes mellitus: molecules of interest and structure-activity relationship. *Curr Med Chem.* 2011;18(17):2661-2672.
47. Miyata Y, Tanaka H, Shimada A, et al. Regulation of adipocytokine secretion and adipocyte hypertrophy by polymethoxyflavonoids, nobiletin and tangeretin. *Life Sci.* . 2011;88(13-14):613-618.
48. MS. W-P. Green tea catechins, caffeine and body-weight regulation. *Physiol Behav.* 2010;100(1):42-46.
49. Hughes L, Arts I, Ambergen T, et al. Higher dietary flavone, flavonol, and catechin intakes are associated with less of an increase in BMI over time in women: a longitudinal analysis from the Netherlands Cohort Study. *Am J Clin Nutr.* 2008;88(5):1341-1352.
50. Nour N. Obesity in resource-poor nations. *Rev Obstet Gynecol.* . 2010;3(4):180-184.
51. Ortega Calvo M, Román Torres P, Lapetra Peralta J. Epistemology as health research propedeutics. *Gac Sanit.* 2011.
52. Lawlor D, Davey Smith G, Ebrahim S. Commentary: the hormone replacement-coronary heart disease conundrum: is this the death of observational epidemiology? *Int J Epidemiol.* 2004;33(3):464-467.
53. Lord S, Gebski V, Keech A. Multiple analyses in clinical trials: sound science or data dredging? *Med J Aust.* . 2004;18(18).
54. Lee J, Williams P, Cheon S. Data mining in genomics. *Clin Lab Med.* 2008;28(1):145-166.

Formatted: Indent: Left: 0", Hanging: 0.5"

Table 1: Sample of the data set presenting the five key health indicators(rate per 100'000) and tea consumption in 8 countries (kg per 100'000)

Formatted: Space After: 0 pt, Line spacing: single, Pattern: Clear (White), Tab stops: 0.64", Left + 1.27", Left + 1.91", Left + 2.54", Left + 3.18", Left + 3.82", Left + 4.45", Left + 5.09", Left + 5.73", Left + 6.36", Left + 7", Left + 7.63", Left + 8.27", Left + 8.91", Left + 9.54", Left + 10.18", Left

Formatted: Font: Not Bold, Font color: Auto, English (U.S.)

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

8/15/2012 6:51:00 PM

8/15/2012 6:51:00 PM

8/15/2012 6:51:00 PM

8/15/2012 6:51:00 PM

8/15/2012 6:51:00 PM

8/15/2012 6:51:00 PM

8/15/2012 6:51:00 PM

8/15/2012 6:51:00 PM

8/15/2012 6:51:00 PM

8/15/2012 6:51:00 PM

8/15/2012 6:51:00 PM

8/15/2012 6:51:00 PM

8/15/2012 6:51:00 PM

8/15/2012 6:51:00 PM

8/15/2012 6:51:00 PM

8/15/2012 6:51:00 PM

8/15/2012 6:51:00 PM

8/15/2012 6:51:00 PM

8/15/2012 6:51:00 PM

8/15/2012 6:51:00 PM

8/15/2012 6:51:00 PM

8/15/2012 6:51:00 PM

8/15/2012 6:51:00 PM

8/15/2012 6:51:00 PM

8/15/2012 6:51:00 PM

8/15/2012 6:51:00 PM

8/15/2012 6:51:00 PM

8/15/2012 6:51:00 PM

8/15/2012 6:51:00 PM

8/15/2012 6:51:00 PM

8/15/2012 6:51:00 PM

8/15/2012 6:51:00 PM

8/15/2012 6:51:00 PM

8/15/2012 6:51:00 PM

8/15/2012 6:51:00 PM

8/15/2012 6:51:00 PM

8/15/2012 6:51:00 PM

8/15/2012 6:51:00 PM

8/15/2012 6:51:00 PM

8/15/2012 6:51:00 PM

8/15/2012 6:51:00 PM

8/15/2012 6:51:00 PM

8/15/2012 6:51:00 PM

8/15/2012 6:51:00 PM

8/15/2012 6:51:00 PM

8/15/2012 6:51:00 PM

8/15/2012 6:51:00 PM

8/15/2012 6:51:00 PM

8/15/2012 6:51:00 PM

8/15/2012 6:51:00 PM

8/15/2012 6:51:00 PM

8/15/2012 6:51:00 PM

8/15/2012 6:51:00 PM

8/15/2012 6:51:00 PM

8/15/2012 6:51:00 PM

8/15/2012 6:51:00 PM

8/15/2012 6:51:00 PM

8/15/2012 6:51:00 PM

8/15/2012 6:51:00 PM

8/15/2012 6:51:00 PM

8/15/2012 6:51:00 PM

8/15/2012 6:51:00 PM

8/15/2012 6:51:00 PM

8/15/2012 6:51:00 PM

8/15/2012 6:51:00 PM

8/15/2012 6:51:00 PM

8/15/2012 6:51:00 PM

8/15/2012 6:51:00 PM

8/15/2012 6:51:00 PM

8/15/2012 6:51:00 PM

8/15/2012 6:51:00 PM

8/15/2012 6:51:00 PM

8/15/2012 6:51:00 PM

8/15/2012 6:51:00 PM

8/15/2012 6:51:00 PM

8/15/2012 6:51:00 PM

8/15/2012 6:51:00 PM

8/15/2012 6:51:00 PM

Figure 1: 2009 Black Tea consumption data in kg/year per inhabitant (source: Euromonitor)

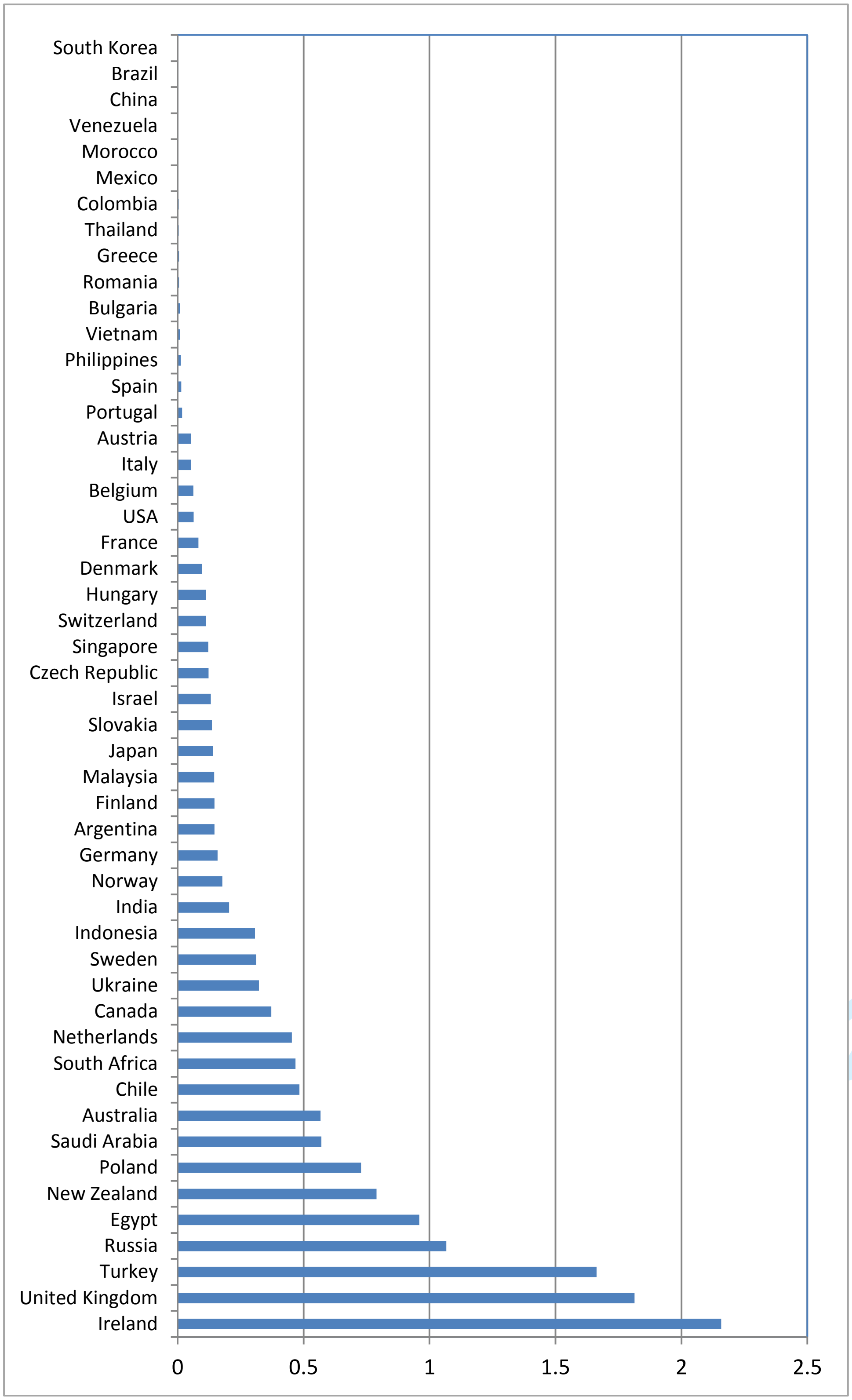
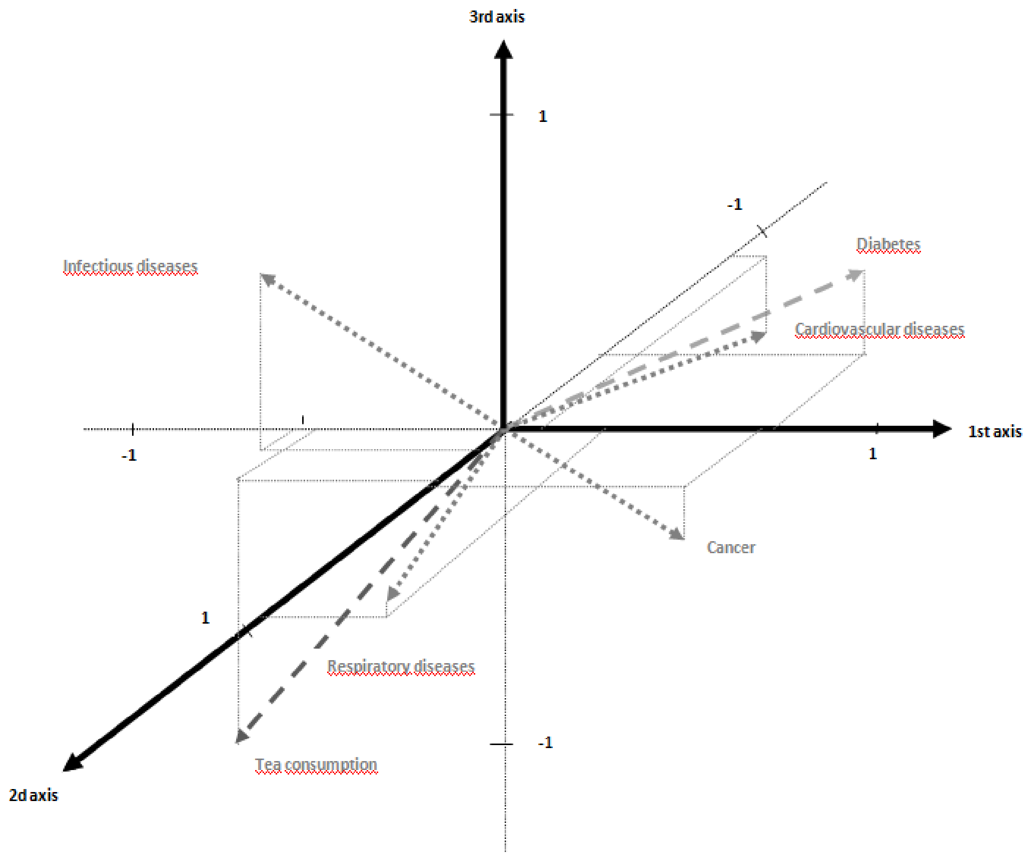


Figure 2: Three dimensional correlation circle of 5 health indicators and BT consumption*



*In this three-dimensional representation, the "infectious disease" vector seems to be close to the BT vector, but is actually represented by a large angle in the third dimension, confirming the poor meaningful correlations between the "infectious diseases" and "BT consumption" variables.

Figure 3: Linear correlation model between black tea consumption (kg per 100,000 inhabitants) and diabetes prevalence (cases per 100,000)