



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

Journal:	<i>BMJ Open</i>
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

SCHOLARONE™
Manuscripts

Peer Review Only

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

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 the BMJ editions and any other BMJGL products and sublicences such use and 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 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 2030¹ 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 diseases². Furthermore, diabetic retinopathy is a common and specific microvascular complication of diabetes, and remains the leading cause of preventable blindness in working-aged people³. With one of the highest prevalences of all human diseases, diabetes is now a global epidemic with devastating health, social and economic consequences⁴. 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 populations⁵. 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 burden⁶. 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 research⁷. 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,

1
2
3 Saudi Arabia, Singapore, Slovakia, South Africa, South Korea, Spain, Sweden, Switzerland, Thailand,
4 Turkey, Ukraine, United Kingdom, USA, Venezuela, Vietnam.

5
6 Epidemiological data were derived from a specific analysis of the World Health Survey (WHS)
7 conducted by the World Health Organization (WHO). Each year, the WHS compiles comprehensive
8 baseline information on the health of populations and health system outcomes⁸. Five key health
9 indicators were selected in 50 countries in both men and women for all age groups: prevalence of
10 respiratory diseases, prevalence of infectious diseases (tuberculosis and HIV), prevalence of cancer,
11 prevalence of cardiovascular diseases and prevalence of diabetes.
12
13
14

15 16 *Methods*

17
18 Data analyses were based on a systematic data-mining approach. Data-mining (sometimes called data or
19 knowledge discovery) is generally defined as the process of analysing data from different perspectives
20 and summarising these data into meaningful information. This approach is useful to analyse data derived
21 from different dimensions or perspectives and to detect potential relationships between variables.
22 Technically, data-mining consists of discovering specific correlations or patterns in large relational
23 databases. Data-mining combines methods from statistics and artificial intelligence with database
24 management and is considered to be an increasingly important tool. It is currently used in a wide range
25 of scientific applications in health⁹⁻¹².
26
27
28
29

30
31 In this study, the data-mining approach used 3 phases: firstly, a "calibrated principal component analyses"
32 (PCA) was used to segment the database composed of 6 variables (BT consumption and the 5 health
33 indicators) into 3 synthetic dimensions; secondly, the 6 variables were represented as vectors in a
34 "correlation circle" to study potential positive or negative correlations; finally, a linear correlation model
35 was tested on selected variables.
36
37
38

39 *Normative principal component analysis (PCA)*

40
41 PCA is a mathematical procedure that uses mathematical projections to convert a set of n possibly
42 correlated variables representing n dimensions into a smaller number of dimensions called "principal
43 components" classically represented in 2 or 3 axes F1, F2, F3. The projections use orthogonal
44 transformations defined in such a way that the first principal component (first axis) has the highest
45 possible variance in order to synthesize most of the initial information. The main objective of PCA is to
46 reduce the dimensionality of the data set. PCA is often presented as a technique of factor analysis for
47 quantitative variables. Multiple Correspondence Analysis (MCA) is another type of factor analysis for
48 quantitative, qualitative and categorical variables and is useful to conduct multi-criteria analyses such as
49 multi-criteria risk assessment¹³. A "normative PCA" was selected for our study, as the 6 variables (BT
50 consumption per capita and 5 key health indicators) are quantitative variables and this analysis was
51 calibrated to study potential correlations.
52
53
54
55
56
57
58
59
60

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: $y = ax + b$ (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

1
2
3 multidimensional table, the overall "quality" (percentage of original variance) of the final projection in 2
4 dimensions was 59% and 74% when projected in 3 dimensions, confirming that the best representation of
5 the dataset should be in 3 dimensions described by axes F1, F2 and F3. The "BT consumption" variable
6 provided a high contributed to the construction on axis F3 (81%). The angle of the vector "BT
7 consumption" with axis F3 was only 22°, confirming the very high contribution of this variable on axis
8 F3. Forty two of the 50 countries were related to this axis F3. The 8 countries not contributing to F3 were
9 Brazil, China, Venezuela, Morocco, Colombia, Viet-Nam, Philippine and Israel, suggesting the absence
10 of any correlation between BT consumption and health indicators in these countries.
11

12
13
14
15 The "correlation circle" (Figure 1) shows that the "BT consumption" vector was strictly opposite the
16 "Diabetes prevalence" vector, establishing a strong statistical negative correlation. Vectors concerning the
17 other key health indicators (infectious diseases, respiratory diseases, cancer and cardiovascular diseases)
18 were represented with a large angle (close to orthogonal) compared to the BT vector, confirming poor
19 statistical relationships between BT and these 4 Health indicators. Of particular interest was the
20 interpretation of the "infectious disease" vector, which seemed to be close to the BT vector in the two
21 dimensional correlation circle, but was actually represented by a large angle in the third dimension. The
22 infectious disease vector was also closer to the centre of the correlation circle, confirming the poor
23 meaningful correlations and potentially hazardous interpretations.
24
25

26
27
28
29
30 The linear correlation model can be expressed as follows:

$$\text{Diabetes prevalence} = a * \text{BT consumption} + b$$

31
32
33
34 Based on 42 countries, the p value of the Fisher-Snedecor test was 0.01, which is highly significant,
35 confirming the relevance of the linear model. The coefficient r^2 was equal to 0.501.
36

37
38 The coefficient $a = -0.0171183$ and a Student's t-test confirmed that this coefficient was significantly
39 different from 0 ($p = 0.001$) with a 5% confidence interval between $[-0.007; -0.027]$. The negativity of the
40 coefficient "a" means that when BT consumption increases, diabetes prevalence decreases, confirming a
41 negative correlation (Figure 2).
42

43
44 Then linear correlation model can be represented by the following formula:

$$\text{Diabetes prevalence} = - 0.0171183 * \text{BT consumption} + 6173.64$$

45 46 47 48 49 50 **Discussion**

51
52 This study establishes, for the first time, a linear statistical relationship between high BT consumption and
53 low diabetes prevalence in the countries that formed the basis for this analysis. As in any database
54 analysis, the very first limitation of this study is related to the quality of the data. WHO prevalence data
55 were obtained from the WHS, which constitutes a convenient and official source of key morbidity
56
57
58
59
60

1
2
3 indicators around the world. The general design of the WHS is based on population sampling organized in
4 the 192 Member States of the United Nations using face-to-face or telephone interviews. As the survey
5 questionnaire offers a menu of choices of modules for each country, and lets the country select the survey
6 approach (Household face-to-face survey, Computer-Assisted Telephone Interview or Computer-Assisted
7 Personal Interview), the quality of data collection can be expected to be heterogeneous around the world.
8 Furthermore, diabetes diagnostic criteria can vary from country to country. On the other hand, any fixed
9 survey design with fixed criteria would not be appropriate everywhere, for example in countries with low
10 telephone network coverage when planning telephone interviews. Other approaches to estimate
11 prevalence of diabetes in the world have been studies using literature and data extrapolations ¹,
12 confirming the growing burden of diabetes.
13
14

15
16
17
18 Another important concern is the interpretation of the established statistical relationship between BT
19 consumption and diabetes prevalence. Using advanced data mining techniques, we tested the potential
20 statistical relationship between BT consumption and 5 health indicators, without any *a priori* assumptions
21 in relation to any of these health indicators. We observed that, among the 5 health indicators, only the
22 "prevalence of diabetes" indicator appeared to have a strong statistical relationship with BT consumption.
23 The relevance and mechanism of this relationship then needs to be discussed. As tea is the most widely
24 used ancient hot beverage in the world, the simple act of putting tea leaves into hot water has provided
25 ancient societies with a tasty beverage with potential medicinal benefits. Two principal varieties of the
26 species are used: the small-leaved Chinese variety (*C. sinensis sinensis*), also used for green tea and white
27 tea, and the large-leaved Assamese variety (*C. sinensis assamica*), which has been traditionally only used
28 for BT. Ancient Chinese civilizations realised that using a special fermentation process, tea leaves would
29 become darker allowing them to be stored for longer periods of time. During this fermentation process, in
30 which green tea oxidises to form black tea, caffeine tends to remain constant, while the types of
31 flavonoids present in the tea differ. Green tea contains simple flavonoids called catechins, whilst BT
32 contains complex flavonoids called theaflavins and thearubigins, which could be the chemical entities
33 responsible for a number of potential health benefits. These tea types were called black tea because of the
34 change in colour of the leaves as a result of this fermentation process. Numerous *in vitro* and *in vivo*
35 studies have demonstrated the health benefits of green tea, mainly in cancer, cardiovascular disease,
36 chronic inflammations or cognitive functions ¹⁴⁻²². However, large-scale clinical dose-effect studies are
37 still missing and it is difficult to interpret the clinical significance of results derived from some
38 biological studies. Considerably fewer studies have been conducted on BT, mostly investigating its
39 antioxidant properties ^{23, 24}, and cardiovascular effects ^{25, 26}. Anti-diabetes properties of BT are suggested
40 in some very specific studies such as a change in pancreatic function in streptozotocin-induced glucose-
41 intolerant rats ^{27, 28}, but also in some human studies together with other hot beverages ²⁹⁻³². Relatively
42 recent interest in BT may be explained by the fact that BT is historically the type of tea most widely
43 consumed in Western countries, probably due to its good storage properties, promoting active trade with
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 tea-producing countries in Asia. Although there has recently been a renewed interest in green tea in
4 industrialized countries due to its popular health benefits, BT represents over ninety percent of all tea sold
5 in the West.
6

7
8 The obesity epidemic in many countries has stimulated interest in food components that may support
9 weight management. Whereas many laboratory and physiological studies have demonstrated the potential
10 effectiveness of BT for the prevention of obesity^{28,29,33,34}, the underlying mechanisms remain unclear.

11 The results of human intervention studies are mixed³⁵ and the role of caffeine has been suggested but not
12 clearly established^{34,36}. Neyestani *et al*³³ found that regular daily intake of BT improves oxidative stress
13 biomarkers and decreases serum C-reactive protein levels in type 2 diabetic patients. Histological studies
14 on pancreas cells published by Manikandan *et al*²⁸ concluded that the BT extract contributes to
15 regeneration of damaged pancreas cells and protects pancreatic beta cells by its antioxidant action.
16

17 Nonetheless, the role of environment, dietary and lifestyle practices is fundamental when comparing
18 health indicators around the world. Psaltopoulou *et al*³⁷ confirmed that low-glycaemic index dietary
19 patterns reduced both fasting blood glucose and glycated proteins independently of carbohydrate
20 consumption. Diets rich in whole-grain, cereal high-fibre products, and non-oil-seed pulses would also be
21 beneficial. As vitamins and minerals play an important role in glucose metabolism, understanding the
22 impact of potential vitamin and mineral deficiencies across cultures is also relevant to better organization
23 of prevention and management of type 2 diabetes^{38,39}. An observational study based on nearly 37,000
24 middle-aged Chinese reported a 14% reduction in the risk of developing type 2 diabetes by drinking one
25 or more cups of tea per day⁴⁰. This was confirmed by two meta-analyses published by Huxley *et al*³⁴ and
26 Jing *et al*⁴¹. Flavonoids are believed to support normal glucose metabolism via anti-inflammatory effects
27 and increased insulin activity^{42,43}. Various studies, especially in Asian populations, confirm that
28 flavonoids present in tea could reduce fat absorption in the gut, would promote fat oxidation in tissues
29 and would increase energy expenditure⁴⁴. An observational study of 4,300 Dutch adults found that
30 flavonoid intake was highest in women who gained the least weight over a 14-year period⁴⁵.

31 Furthermore, as physical activity with or without diet contributes to a healthier lifestyle, this important
32 factor must be considered when comparing health indicators between industrialized and emerging
33 countries. Given rapid population growth, increased urbanization, and adverse lifestyle changes, the
34 obesity/type 2 diabetes epidemic in resource-poor nations was predicted in the 1990s and has now been
35 fully confirmed⁴⁶, underlying the importance of a better understanding of predictive and potentially
36 protective factors.
37

38 The number of factors contributing to the growth of diabetes and obesity in the world confirm that
39 "correlation does not imply causality", and that a significant linear correlation between BT consumption
40 and diabetes prevalence does not imply that low BT consumption could *cause* diabetes. If one factor is
41 established as causing another, then the two factors are most certainly correlated. However, the opposite
42
43
44
45
46
47
48
49
50

1
2
3 cannot be concluded. Thus, a correlation can only indicate a potential direct or indirect possible cause,
4 which then needs to be further investigated. This paradigm and the connotations of causality may be the
5 most important considerations affecting biostatistics in major epidemiological study designs⁴⁷. A well
6 known example of epidemiological cause-and-effect misinterpretations is the correlation that was
7 established between hormone replacement therapy and a lower incidence of coronary heart disease. This
8 association has been more recently explained by the fact that women taking hormone replacement therapy
9 were more likely to come from higher socio-economic levels, which could explain the lower incidence of
10 coronary heart disease⁴⁸. Establishing causality is one of the most difficult challenges in public health.
11 For instance, in clinical research, randomized controlled clinical trials are performed to establish potential
12 significant differences between two groups. However, establishing a difference is not a demonstration of
13 causality. Another example is case-control studies, which compare individuals with a specific disease
14 ("cases") with a group of individuals without the disease ("controls"). An association between the
15 hypothesized exposure and the disease studied would be reflected by a higher proportion in exposed
16 cases, but this cannot constitute a real demonstration of causality. A potential causality can only be
17 established with the convergence of interdisciplinary scientific evidence (biological, physiological,
18 epidemiological, etc.) and reasonable explanations based on longitudinal studies. In our study, biological,
19 physiological and some epidemiological studies can be considered to provide evidence linking BT
20 consumption of BT and the prevalence of diabetes. However, a large-scale, longitudinal, prospective
21 case-control study comparing high BT consumption versus no consumption and diabetes prevalence
22 would be useful to confirm these findings.
23
24
25
26
27
28
29
30
31
32

33
34 Beyond the causality issue, a frequent criticism of using data-mining was based on the confusion between
35 *data-mining* and *data-dredging* techniques. While a data-mining approach is based on searching for
36 combinations of variables that might show potential correlations, data-dredging (also called "data-
37 fishing") can generate misleading results⁴⁹. When a number of hypotheses are tested, it is expected that
38 some will falsely appear to be statistically significant, since every database can contain potential random
39 correlations. A robust data mining approach must therefore always be based on a clear research strategy
40 and a limited number of relevant meaningful assumptions. In our assessment, we used a systematic data
41 mining approach to test potential correlations between 6 selected variables (BT consumption and 5 key
42 health indicators). PCA was used to describe and structure the dataset before testing any correlations. In
43 our study, only one linear correlation model was constructed between BT consumption and diabetes
44 prevalence, based on the most relevant association suggested by the PCA. This consistent approach is
45 quite different from screening numerous cross-regression analyses between all variables of one particular
46 dataset. The data-mining approach can be considered to be a "radar tracking system", allowing detection,
47 tracking and classification of potential "targets" in the framework of a particular environment. This is
48 particularly useful when exploring complex databases, as data-mining can identify original statistical
49 evidence, which would never be discovered by means of classical statistical techniques. As an example,
50
51
52
53
54
55
56
57
58
59
60

the significant progress in genomics would not have been possible without the use of data-mining techniques⁵⁰. 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.

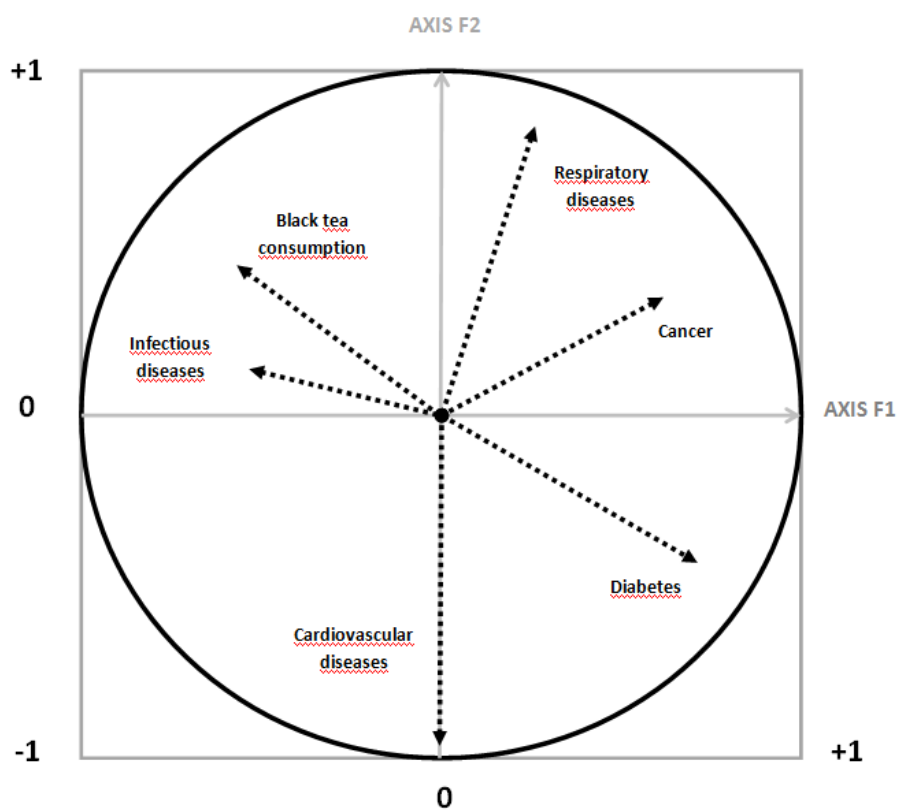
References

1. Shaw JE, Sicree RA, Zimmet PZ. Global estimates of the prevalence of diabetes for 2010 and 2030. *Diabetes Research and Clinical Practice*. 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 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, Puanggraphant 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. Pełal A, Drózdź 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. Psaltopoulou 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.

- 1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
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.

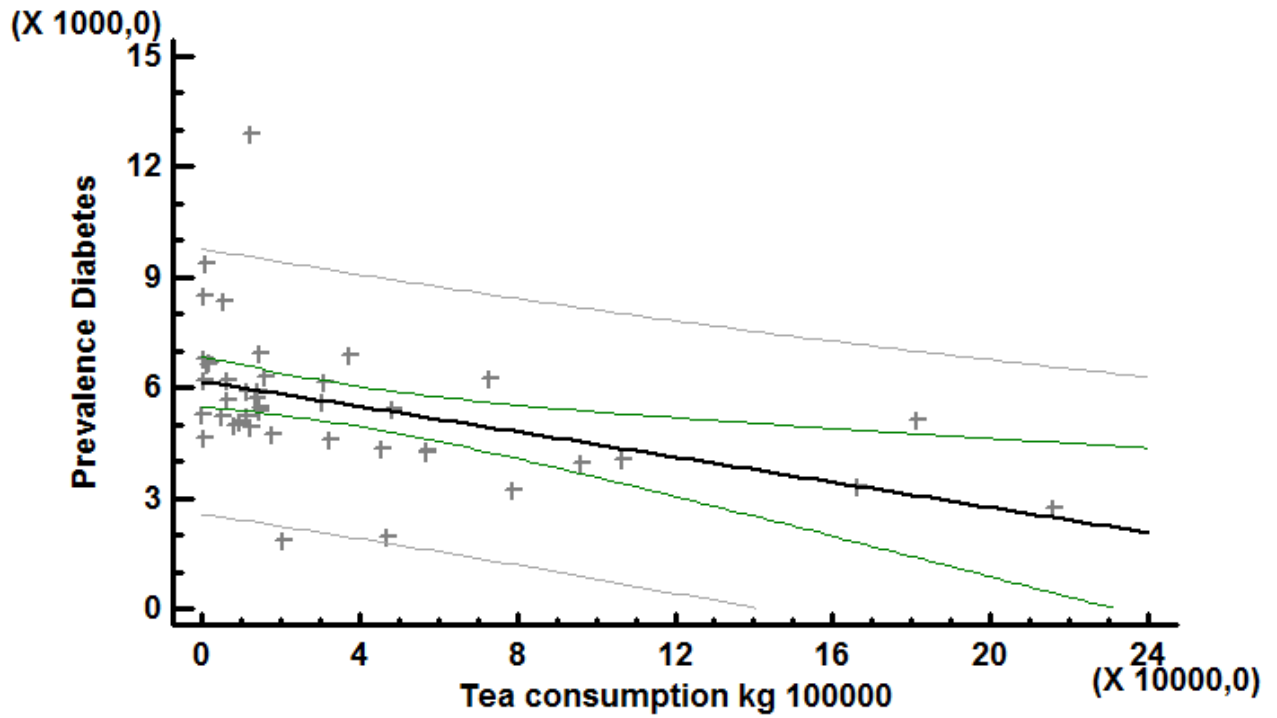
Figure 1: Two dimensional Correlation circle of the 5 Health indicators and BT consumption*



*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.

View only

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:	<i>BMJ Open</i>
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

SCHOLARONE™
Manuscripts

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 BMJPG 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'Aéroport, 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 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

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 2030¹ 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 diseases². Furthermore, diabetic retinopathy is a common and specific microvascular complication of diabetes, and remains the leading cause of preventable blindness in working-aged people³. With one of the highest prevalences of all human diseases, diabetes is now a global epidemic with devastating health, social and economic consequences⁴. 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 populations⁵. 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 burden⁶. 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 research⁹. 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

1
2
3 Zealand, Norway, Philippines, Poland, Portugal, Romania, Russia, Saudi Arabia, Singapore, Slovakia,
4 South Africa, South Korea, Spain, Sweden, Switzerland, Thailand, Turkey, Ukraine, United Kingdom,
5 USA, Venezuela, Vietnam (Figure 1). Highest BT consumptions (kg/year per inhabitant) are observed in
6 Ireland (2.1576), UK (1.8137), Turkey (1.6631) and Russia (1.0668). Lowest BT consumptions are
7 observed in South Korea (0.0007), Brazil (0.001) and China (0.0011), as the Chinese population drinks 30
8 times more green tea (0.036 kg per inhabitant) than black tea.

9
10
11 Epidemiological data were derived from a specific analysis of the World Health Survey (WHS)
12 conducted by the World Health Organization (WHO). Each year, the WHS compiles comprehensive
13 baseline information on the health of populations and health system outcomes¹⁰. Using the 2009 dataset,
14 five key health indicators were selected in 50 countries in both men and women for all age groups:
15 prevalence of respiratory diseases, prevalence of infectious diseases (tuberculosis and HIV), prevalence
16 of cancer, prevalence of cardiovascular diseases and prevalence of diabetes (Figure 2).
17
18
19
20

21 *Methods*

22
23
24 Data analyses were based on a systematic data mining approach. Data mining (sometimes called data or
25 knowledge discovery) is generally defined as the process of analysing data from different perspectives
26 and summarising these data into meaningful information. This approach is useful to analyse data derived
27 from different dimensions or perspectives and to detect potential relationships between variables.
28 Technically, data mining consists of discovering specific correlations or patterns in large relational
29 databases. Data mining combines methods from statistics and artificial intelligence with database
30 management and is considered to be an increasingly important tool. It is currently used in a wide range
31 of scientific applications in health¹¹⁻¹⁴.
32
33
34
35

36 In this study, the data mining approach used 3 phases: firstly, a "calibrated principal component analyses"
37 (PCA) was used to segment the database composed of 6 variables (BT consumption and the 5 health
38 indicators) into 3 synthetic dimensions; secondly, the 6 variables were represented as vectors in a
39 "correlation circle" to study potential positive or negative correlations; finally, a linear correlation model
40 was tested on selected variables.
41
42
43

44 *Normative principal component analysis (PCA)*

45
46 PCA is a mathematical procedure that uses mathematical projections to convert a set of n possibly
47 correlated variables representing n dimensions into a smaller number of dimensions called "principal
48 components" classically represented in 2 or 3 axes F1, F2, F3. The projections use orthogonal
49 transformations defined in such a way that the first principal component (first axis) has the highest
50 possible variance in order to synthesize most of the initial information. The main objective of PCA is to
51 reduce the dimensionality of the data set. PCA is often presented as a technique of factor analysis for
52 quantitative variables. Multiple Correspondence Analysis (MCA) is another type of factor analysis for
53
54
55
56
57
58
59
60

1
2
3 quantitative, qualitative and categorical variables and is useful to conduct multi-criteria analyses such as
4 multi-criteria risk assessment¹⁵. A "normative PCA" was selected for our study, as the 6 variables (BT
5 consumption per capita and 5 key health indicators) are quantitative variables and this analysis was
6 calibrated to study potential correlations.
7
8

9 10 *Correlation circle*

11 The correlation circle shows a projection of the initial variables in a dimensional space represented by
12 axes F1 and F2¹⁶. Variables are presented as vectors from the centre. When two vectors are close to the
13 correlation circle, they can be: i) close to each other, meaning a positive correlation ii) orthogonal from
14 each other, meaning that they are not correlated iii) on the opposite side from the centre, meaning a
15 significant negative correlation. When some vectors are close to the centre, this means that some
16 information is carried on other axes, and that any interpretation might be hazardous. This can be
17 confirmed by looking at another correlation circle constructed with axes F1 and F3 or with axes F2 and
18 F3. The correlation circle is then used to identify the potential proximity with the 6 vectors and to assess
19 their potential correlations. Should a vector representing the variable "BT consumption" be close to the
20 correlation circle and point to a similar direction compared to any of the other 5 vectors representing
21 health indicators, this would indicate a positive correlation between the two variables.
22
23
24
25
26
27
28
29

30 31 *Linear correlation model*

32 Once identified by the correlation circle, potential correlations between BT consumption and one or more
33 health indicators can be described using key statistical parameters, such as the coefficient r^2 and the
34 statistical significance p . Using a linear correlation model between BT consumption and one health
35 indicator then determines the extent to which the values of these two variables are potentially
36 "proportional" to each other (BT consumption increases or decreases with one specific disease
37 prevalence). The linear model formula is: $y = ax + b$ (y = health indicator; x = BT consumption; a and b
38 are the model calculated coefficients). The structure of this formula suggests that the variables x and y are
39 linearly related and thus proportional; that is, the correlation is high if it can be represented by a straight
40 line (upwards or downwards slope). If so, this line will represent the linear model, also called a
41 "regression line" or "least squares line" because the sum of the squared distances of all the data points
42 from the line is the lowest possible. The coefficient r^2 (coefficient of determination) represents the
43 proportion of common variations between the two variables and establishes the "strength" of the
44 relationship. In order to evaluate the potential correlation between BT consumption and one specific
45 health indicator, it is therefore important to know r^2 , the statistical significance p of the correlation
46 (calculated by a Fisher-Snedecor test) and the statistical significance of the difference to 0 of the
47 coefficient " a " (Student's t-test).
48
49
50
51
52
53
54
55
56
57
58
59
60

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:

$$\text{Diabetes prevalence} = a * \text{BT consumption} + b$$

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).

1
2
3 Then linear correlation model can be represented by the following formula and is presented in Figure 4::
4

$$5 \quad \text{Diabetes prevalence} = - 0.0171183 * \text{BT consumption} + 6173.64$$

6
7

8 9 Discussion

10 This study establishes, for the first time, a linear statistical relationship between high BT consumption and
11 low diabetes prevalence in the countries that formed the basis for this analysis. As in any database
12 analysis, the very first limitation of this study is related to the quality of the data. WHO prevalence data
13 were obtained from the WHS, which constitutes a convenient and official source of key morbidity
14 indicators around the world. The general design of the WHS is based on population sampling organized in
15 the 192 Member States of the United Nations using face-to-face or telephone interviews. As the survey
16 questionnaire offers a menu of choices of modules for each country, and lets the country select the survey
17 approach (Household face-to-face survey, Computer-Assisted Telephone Interview or Computer-Assisted
18 Personal Interview), the quality of data collection can be expected to be heterogeneous around the world.
19

20 Furthermore, some of the selected health indicators represent a group of diseases, such as infectious
21 diseases (tuberculosis and HIV) and cancer. The heterogeneity of these indicators can make it difficult to
22 establish any potential statistical relationships. Although more homogeneous, health indicators such as
23 diabetes depend on diagnostic criteria, which can vary from country to country. On the other hand, any
24 fixed survey design with fixed criteria would not be appropriate everywhere, for example in countries
25 with low telephone network coverage when planning telephone interviews. Other approaches to estimate
26 prevalence of diabetes in the world have been studies using literature and data extrapolations ¹,
27 confirming the growing burden of diabetes.
28

29 Another important concern is the interpretation of the established statistical relationship between BT
30 consumption and diabetes prevalence. Using advanced data mining techniques, we tested the potential
31 statistical relationship between BT consumption and 5 health indicators, without any *a priori* assumptions
32 in relation to any of these health indicators. We observed that, among the 5 health indicators, only the
33 "prevalence of diabetes" indicator appeared to have a strong statistical relationship with BT consumption.
34 The proposed epidemiological approach considers the population as the unit of analysis rather than an
35 individual and can be presented as an ecological study, which is considered to be inferior to case-control
36 studies in the context of evidence-based medicine. In an ecological study, no information is available
37 about the individual members of the populations compared, whereas in a case-control study, information
38 is reported for each individual. However, ecological studies can be very useful for international
39 comparisons, while case-control studies are exclusively based on local information. Furthermore, when
40 strong correlations have been established, the results of ecological studies can suggest further evidence-
41 based studies, investigating the relevance and mechanism of the statistical relationship. Various study
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 designs have already been used to assess the potential benefits of tea. As this is the most widely used
4 ancient hot beverage in the world, the simple act of putting tea leaves into hot water has provided ancient
5 societies with a tasty beverage associated with the observation of certain medicinal benefits. Two
6 principal varieties of the species are used: the small-leaved Chinese variety (*C. sinensis sinensis*), also
7 used for green tea and white tea, and the large-leaved Assamese variety (*C. sinensis assamica*), which has
8 been traditionally used only for BT. Ancient Chinese civilizations realised that using a special
9 fermentation process, tea leaves would become darker allowing them to be stored for longer periods of
10 time. During this fermentation process, in which green tea oxidises to form black tea, caffeine tends to
11 remain constant, while the types of flavonoids present in the tea differ. Green tea contains simple
12 flavonoids called catechins, whilst BT contains complex flavonoids called theaflavins and thearubigins,
13 which could be the chemical entities responsible for a number of potential health benefits. These tea types
14 were called black tea because of the change in colour of the leaves as a result of this fermentation process.
15 Numerous *in vitro* and *in vivo* studies have demonstrated the health benefits of green tea, mainly in
16 cancer, cardiovascular disease, chronic inflammation or cognitive functions¹⁷⁻²⁵. However, large-scale
17 clinical dose-effect studies are still missing and it is difficult to interpret the clinical significance of results
18 derived from some biological studies. Considerably fewer studies have been conducted on BT, mostly
19 investigating its antioxidant properties^{26,27}, and cardiovascular effects^{28,29}. Anti-diabetes properties of
20 BT have been suggested by several very specific studies, such as a change in pancreatic function in
21 streptozotocin-induced glucose-intolerant rats^{30,31}, but also in some human studies also investigating
22 other hot beverages³²⁻³⁵. The relatively recent interest in BT may be explained by the fact that BT is
23 historically the type of tea most widely consumed in Western countries, probably due to its good storage
24 properties, promoting active trade with tea-producing countries in Asia. Although there has recently been
25 a renewed interest in green tea in industrialized countries due to its popular health benefits, BT represents
26 over ninety percent of all tea sold in the West.

27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

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 Singapore³⁶.

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 mixed³⁹ 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

1
2
3 practices is fundamental when comparing health indicators around the world. Psaltopoulou *et al*⁴¹
4 confirmed that low-glycaemic index dietary patterns reduced both fasting blood glucose and glycosylated
5 proteins independently of carbohydrate consumption. Diets rich in whole-grain, cereal high-fibre
6 products, and non-oil-seed pulses would also be beneficial. As vitamins and minerals play an important
7 role in glucose metabolism, understanding the impact of potential vitamin and mineral deficiencies across
8 cultures is also relevant to better organization of prevention and management of type 2 diabetes^{42,43}. An
9 observational study based on nearly 37,000 middle-aged Chinese reported a 14% reduction in the risk of
10 developing type 2 diabetes by drinking one or more cups of tea per day⁴⁴. This was confirmed by two
11 meta-analyses published by Huxley *et al*³⁸ and Jing *et al*⁴⁵. Flavonoids are believed to support normal
12 glucose metabolism via anti-inflammatory effects and increased insulin activity^{46,47}. Various studies,
13 especially in Asian populations, confirm that flavonoids present in green tea could reduce fat absorption
14 in the gut, may promote fat oxidation in tissues and may increase energy expenditure⁴⁸. An observational
15 study of 4,300 Dutch adults found that flavonoid intake was highest in women who gained the least
16 weight over a 14-year period⁴⁹. Furthermore, as physical activity with or without diet contributes to a
17 healthier lifestyle, this important factor must be considered when comparing health indicators between
18 industrialized and emerging countries. Given rapid population growth, increased urbanization, and
19 adverse lifestyle changes, the obesity/type 2 diabetes epidemic in resource-poor nations was predicted in
20 the 1990s and has now been fully confirmed⁵⁰, underlying the importance of a better understanding of
21 predictive and potentially protective factors.
22
23

24
25
26 The number of factors contributing to the growth of diabetes and obesity in the world confirm that
27 "correlation does not imply causality", and that a significant linear correlation between BT consumption
28 and diabetes prevalence does not imply that low BT consumption could *cause* diabetes. If one factor is
29 established as causing another, then the two factors are most certainly correlated. However, the opposite
30 cannot be concluded. Thus, a correlation can only indicate a potential direct or indirect possible cause,
31 which then needs to be further investigated. This paradigm and the connotations of causality may be the
32 most important considerations affecting biostatistics in major epidemiological study designs⁵¹. A well
33 known example of epidemiological cause-and-effect misinterpretations is the correlation that was
34 established between hormone replacement therapy and a lower incidence of coronary heart disease. This
35 association has been more recently explained by the fact that women taking hormone replacement therapy
36 were more likely to come from higher socio-economic levels, which could explain the lower incidence of
37 coronary heart disease⁵². Establishing causality is one of the most difficult challenges in public health.
38 For instance, in clinical research, randomized controlled clinical trials are performed to establish potential
39 significant differences between two groups. However, establishing a difference is not a demonstration of
40 causality. Another example is case-control studies, which compare individuals with a specific disease
41 ("cases") with a group of individuals without the disease ("controls"). An association between the
42 hypothesized exposure and the disease studied would be reflected by a higher proportion in exposed
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 cases, but this cannot constitute a real demonstration of causality. A potential causality can only be
4 established with the convergence of interdisciplinary scientific evidence (biological, physiological,
5 epidemiological, etc.) and reasonable explanations based on longitudinal studies.
6

7
8 Ecological research can address important issues that cannot be easily addressed by other study designs.
9 They are frequently used where alternative study designs are not possible (eg, randomized control trials),
10 such as when investigating the effect of geographical factors on disease incidence. Our approach to BT
11 consumption presents a number of limitations like all ecological studies because factors other than dietary
12 habits may be the most important determinants of variations in diabetes prevalence across communities.
13 For example, it is possible that other unmeasured confounding factors (eg, genetic differences) may
14 explain some of the observed regional variations. Due to the large number of potential determinants of
15 diabetes prevalence, including patient-, physician-, hospital-, and community-related variables, it is
16 difficult to identify with certainty all of the causes of the regional variations of diabetes prevalence, and
17 additional follow-up studies should be considered to confirm the hypotheses generated by this type of
18 study.
19

20
21 A number of biological, physiological and epidemiological studies have provided evidence linking BT
22 consumption and glucose metabolism^{26, 30, 31, 37-39, 46, 47}. However, a large-scale, longitudinal, prospective
23 case-control study comparing high BT consumption versus no consumption and diabetes prevalence
24 would be useful to confirm these findings.
25

26
27 Beyond the causality issue, a frequent criticism of using data mining was based on the confusion between
28 *data mining* and *data dredging* techniques. While a data mining approach is based on searching for
29 combinations of variables that might show potential correlations, data-dredging (also called "data
30 fishing") can generate misleading results⁵³. When a number of hypotheses are tested, it is expected that
31 some will falsely appear to be statistically significant, since every database can contain potential random
32 correlations. A robust data mining approach must therefore always be based on a clear research strategy
33 and a limited number of relevant meaningful assumptions. In our assessment, we used a systematic data
34 mining approach to test potential correlations between 6 selected variables (BT consumption and 5 key
35 health indicators). PCA was used to describe and structure the dataset before testing any correlations. In
36 our study, only one linear correlation model was constructed between BT consumption and diabetes
37 prevalence, based on the most relevant association suggested by the PCA. This consistent approach is
38 quite different from screening numerous cross-regression analyses between all variables of one particular
39 dataset. The data mining approach can be considered to be a "radar tracking system", allowing detection,
40 tracking and classification of potential "targets" in the framework of a particular environment. This is
41 particularly useful when exploring complex databases, as data mining can identify original statistical
42 evidence, which would never be discovered by means of classical statistical techniques. As an example,
43 the significant progress in genomics would not have been possible without the use of data mining
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

techniques⁵⁴. 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 Practice*. 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.

- 1
 - 2
 - 3
 - 4
 - 5
 - 6
 - 7
 - 8
 - 9
 - 10
 - 11
 - 12
 - 13
 - 14
 - 15
 - 16
 - 17
 - 18
 - 19
 - 20
 - 21
 - 22
 - 23
 - 24
 - 25
 - 26
 - 27
 - 28
 - 29
 - 30
 - 31
 - 32
 - 33
 - 34
 - 35
 - 36
 - 37
 - 38
 - 39
 - 40
 - 41
 - 42
 - 43
 - 44
 - 45
 - 46
 - 47
 - 48
 - 49
 - 50
 - 51
 - 52
 - 53
 - 54
 - 55
 - 56
 - 57
 - 58
 - 59
 - 60
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. Pełal A, Drózdź 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. Psaltopoulou 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.

- 1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
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.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

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

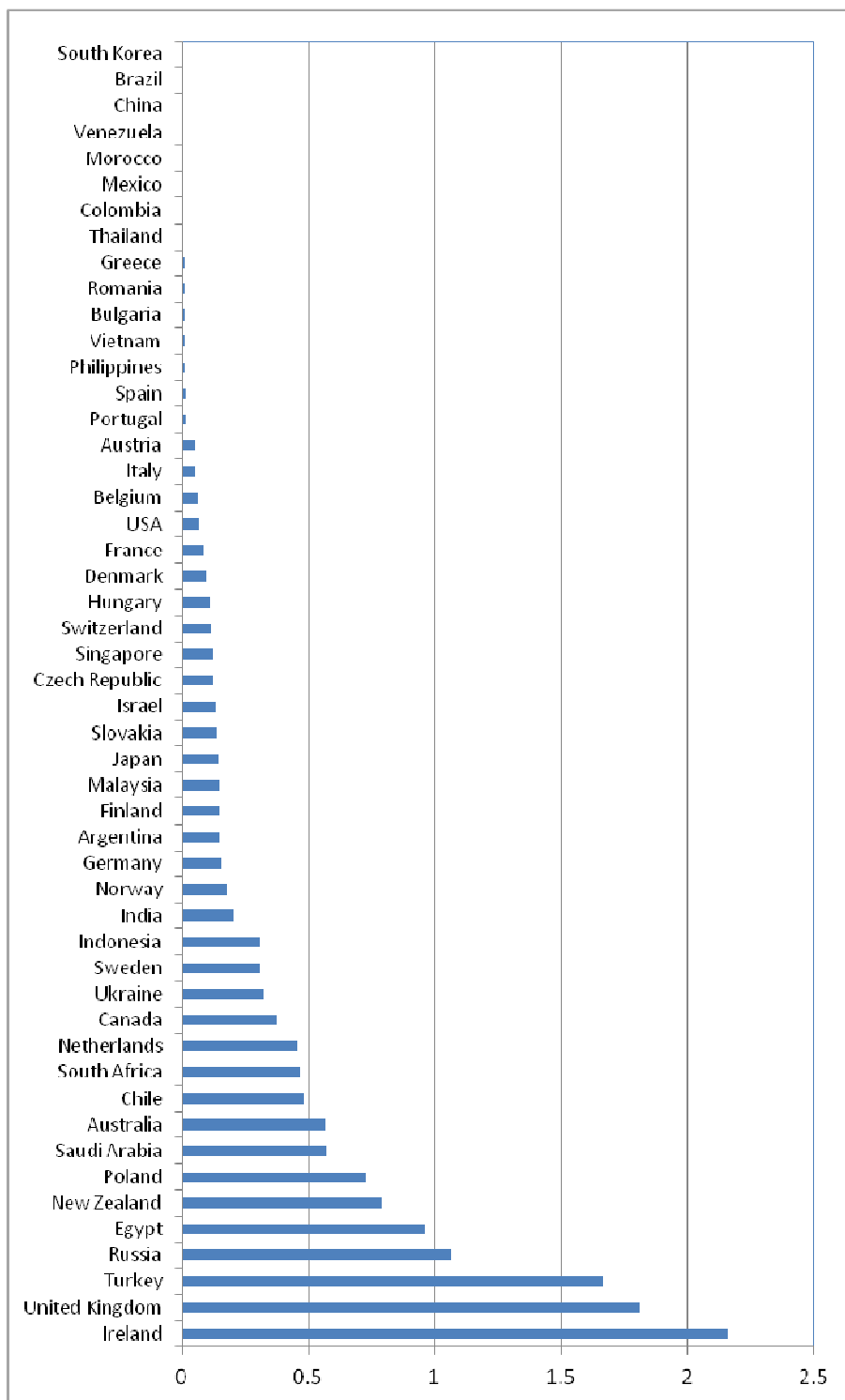


Figure 2: 2009 prevalence (per 100,000) of key health indicators (source: WHO)

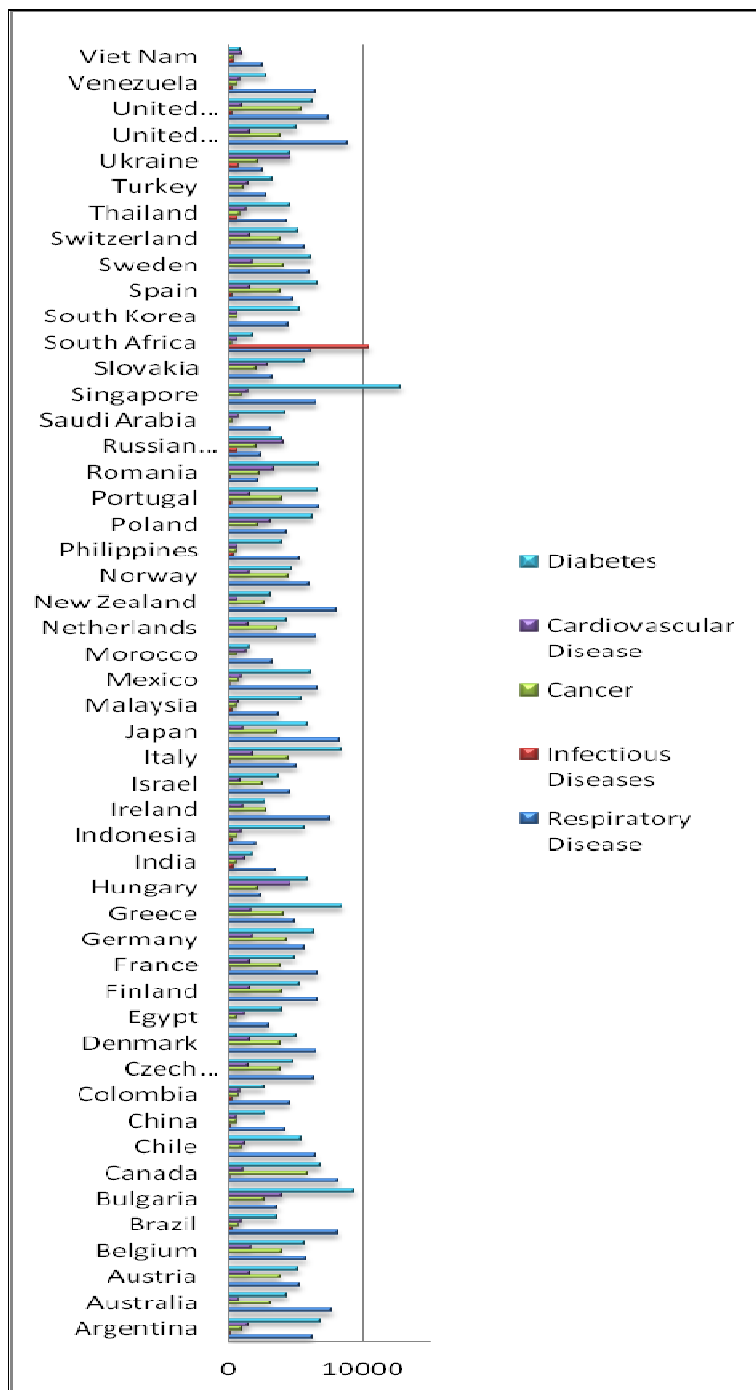
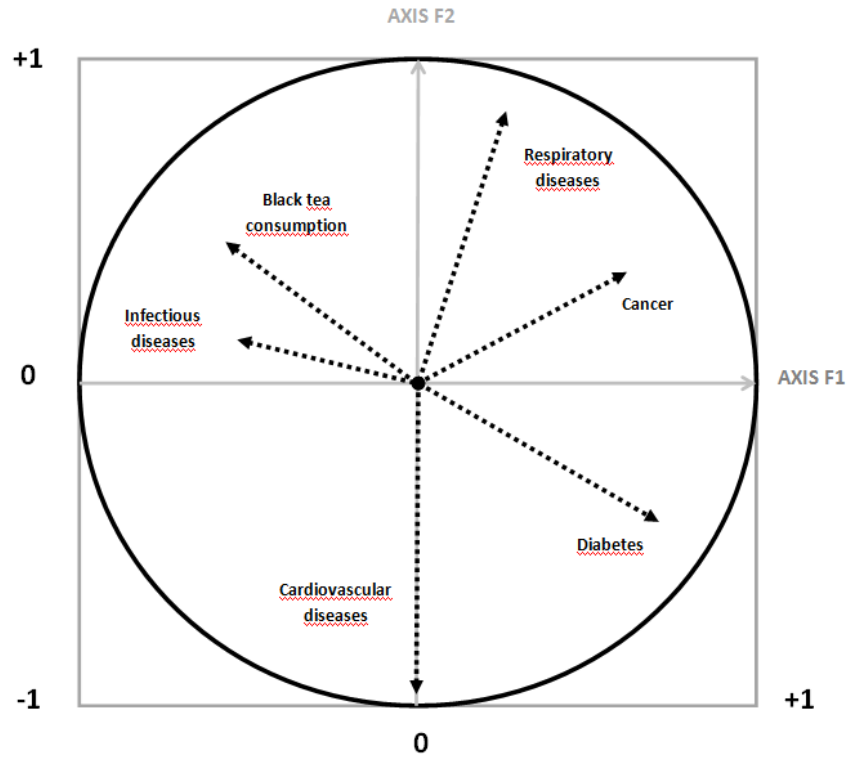
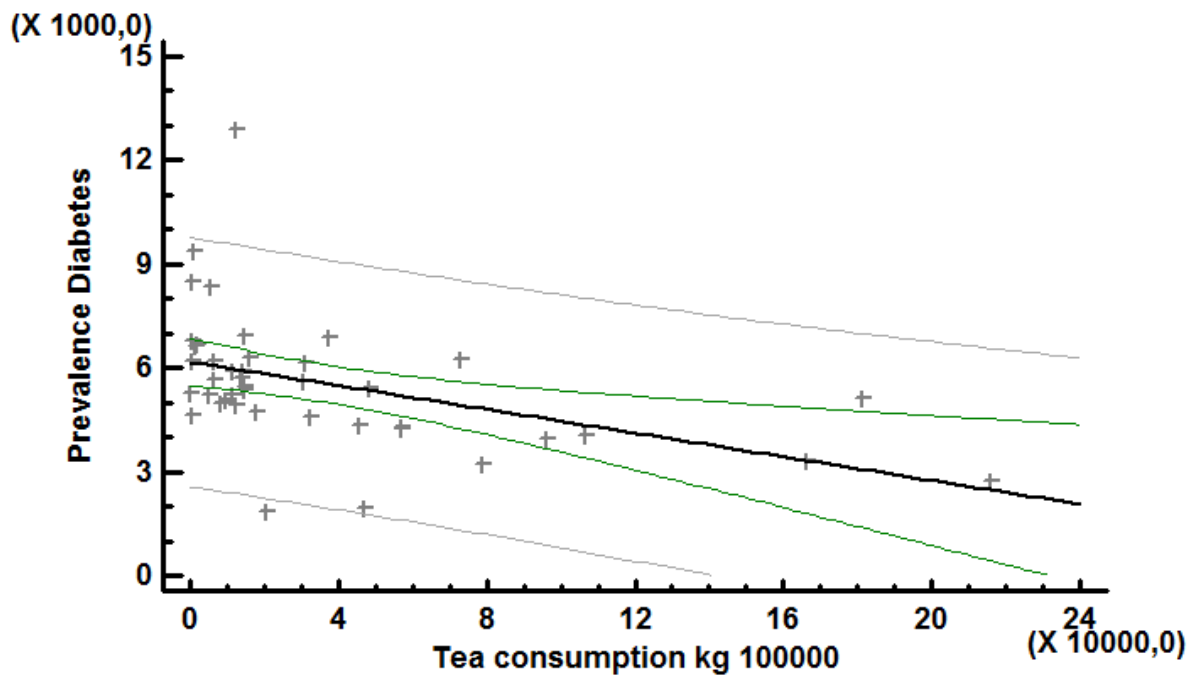


Figure 3: Two dimensional correlation circle of 5 health indicators and BT consumption*



**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:	<i>BMJ Open</i>
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

SCHOLARONE™
Manuscripts

Peer Review Only

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

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 BMJPG 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'Aéroport, 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.

1
2
3 - Diabetes prevalence data were obtained from the World Health Survey implemented by the World
4 Health Organization, which constitutes an official source of key morbidity indicators around the world.
5 However, the quality of data collection can be expected to be heterogeneous around the world and
6 diabetes diagnostic criteria can vary from country to country.

7
8 - Another important concern is the interpretation of the established statistical relationship between BT
9 consumption and diabetes prevalence. The numerous factors contributing to the growth of diabetes and
10 obesity throughout the world confirm that "correlation does not imply causality", and that a significant
11 linear correlation between BT consumption and diabetes prevalence does not imply that low BT
12 consumption could cause diabetes. A correlation can only indicate a potential direct or indirect cause,
13 which then needs to be further investigated.

14
15 - A frequent criticism of the use of data mining is based on the confusion between data mining and data
16 dredging techniques. While a data mining approach is based on searching for combinations of variables
17 that might show potential correlations, data dredging can generate misleading results. When a number
18 of hypotheses are tested, it is expected that some will falsely appear to be statistically significant, since
19 every database can contain potential random correlations. A robust data mining approach must
20 therefore always be based on a clear research strategy and a limited number of relevant meaningful
21 assumptions.
22

23
24 -A classical criticism of this approach is the "ecological fallacy", corresponding to a logical fallacy in
25 interpretation of the observed correlations at the population level, assuming that they can be applied at
26 the individual level. Our study on black tea does not comprise any potential logical fallacy, as it was not
27 used as the basis for any individual assumptions.
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

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* experiments¹⁻⁹. 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 disease¹⁰, 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 actions¹¹ 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 effects¹³. Moreover, dietary consumption of tea catechins would have beneficial effects on blood pressure and lipid parameters¹⁴. 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 cancer¹⁵. Anti-diabetes properties of BT have been suggested by several very specific studies, such as a change in pancreatic function in streptozotocin-

1
2
3 induced glucose-intolerant rats ^{16,17}, but also in some human studies investigating other hot beverages ¹⁸⁻
4 ²¹. The relatively recent interest in BT may be explained by the fact that BT is historically the type of tea
5 most widely consumed in Western countries, probably due to its good storage properties, promoting
6 active trade with tea-producing countries in Asia. Although there has recently been a renewed interest in
7 green tea in industrialized countries, BT represents over ninety percent of all tea sold in the West. Despite
8 the number of publications investigating the effects of tea components and green tea in particular, large-
9 scale clinical dose-effect studies are still lacking and it is difficult to interpret the clinical significance of
10 results derived from some biological studies. Considerably fewer studies have been conducted
11 specifically on BT, mostly investigating its antioxidant properties ^{22,23}, and cardiovascular effects ^{24,25}. It
12 could therefore be interesting to investigate some key dietary habits in relation with lifestyle and health
13 effects at a global level, in view of the perceived positive health effects of black tea (BT), which have
14 been described for centuries ^{26,27}. Because of the complexity of implementing international prospective
15 studies and the difficulty of conducting meta-analyses on a large number of heterogeneous local studies,
16 potential correlations between BT consumption and epidemiological data around the world could be
17 investigated by using advanced exploratory statistical methods. The objective of this original research
18 was to investigate potential statistical relationships between BT consumption and the following five key
19 health indicators: respiratory diseases, infectious diseases, cancer, cardiovascular diseases and diabetes.
20
21
22
23
24
25
26
27
28
29
30

31 **Material and Method**

32 *Data sources*

33
34
35 BT consumption data were derived from a specific international trade survey compiling sales data
36 conducted in 2009 by Euromonitor International, an independent agency specialized in market research ²⁸.
37 Consumption data are derived from black tea international trading registries, used by black tea importers
38 to adapt international orders to local sales. Yearly consumption data expressed in kilograms per capita
39 were available for the following 50 countries: Argentina, Australia, Austria, Belgium, Brazil, Bulgaria,
40 Canada, Chile, China, Colombia, Czech Republic, Denmark, Egypt, Finland, France, Germany, Greece,
41 Hungary, India, Indonesia, Ireland, Israel, Italy, Japan, Malaysia, Mexico, Morocco, Netherlands, New
42 Zealand, Norway, Philippines, Poland, Portugal, Romania, Russia, Saudi Arabia, Singapore, Slovakia,
43 South Africa, South Korea, Spain, Sweden, Switzerland, Thailand, Turkey, Ukraine, United Kingdom,
44 USA, Venezuela, Vietnam (Figure 1). Highest BT consumptions (kg/year per inhabitant) are observed in
45 Ireland (2.1576), UK (1.8137), Turkey (1.6631) and Russia (1.0668). Lowest BT consumptions are
46 observed in South Korea (0.0007), Brazil (0.001) and China (0.0011), as the Chinese population drinks 30
47 times more green tea (0.036 kg per inhabitant) than black tea. Epidemiological data were derived from a
48 specific extraction from the World Health Survey (WHS) conducted by the World Health Organization
49 (WHO). Each year, the WHS compiles comprehensive baseline information on the health of populations
50
51
52
53
54
55
56
57
58
59
60

1
2
3 and health system outcomes²⁹. Using the 2009 dataset (sample presented in Table 1), five key health
4 indicators were selected in 50 countries in both men and women for all age groups: prevalence of
5 respiratory diseases, prevalence of infectious diseases (tuberculosis and HIV), prevalence of cancer,
6 prevalence of cardiovascular diseases and prevalence of diabetes.
7
8

9 *Methods*

10
11 This ecological study used a data mining approach structured in 3 phases: firstly, a "calibrated principal
12 component analysis" (PCA) was used to segment the database composed of 6 variables (BT consumption
13 and the 5 health indicators) into 3 synthetic dimensions represented by 3 axes which can be considered as
14 the mathematical projection of the 6 dimensions defined by the 6 variables into 3 dimensions; secondly,
15 the 6 variables were represented as vectors in a "correlation circle" to study potential positive or negative
16 correlations; finally, a linear correlation model was tested on selected variables.
17
18
19

20 *Normative principal component analysis (PCA)*

21
22 PCA is a mathematical procedure that uses mathematical projections to convert a set of n possibly
23 correlated variables representing n dimensions into a smaller number of dimensions called "principal
24 components" classically represented in 2 or 3 axes F1, F2, F3. The projections use orthogonal
25 transformations defined in such a way that the first principal component (first axis) has the highest
26 possible variance in order to synthesize most of the initial information. The main objective of PCA is to
27 reduce the dimensionality of the dataset. PCA is often presented as a technique of factor analysis for
28 quantitative variables. Multiple Correspondence Analysis (MCA) is another type of factor analysis for
29 quantitative, qualitative and categorical variables and is useful to conduct multi-criteria analyses such as
30 multi-criteria risk assessment³⁰. A "normative PCA" was selected for our study, as the 6 variables (BT
31 consumption per capita and 5 key health indicators) are quantitative variables and this analysis was
32 calibrated to study potential correlations.
33
34
35
36
37
38
39

40 *Correlation circle*

41
42 The correlation circle shows a projection of the initial variables in a dimensional space which can be
43 represented in two or three dimensions³¹. Variables are presented as vectors from the centre. When two
44 vectors are close to the correlation circle, they can be: i) close to each other, meaning a positive
45 correlation ii) orthogonal from each other, meaning that they are not correlated iii) on the opposite side
46 from the centre, meaning a significant negative correlation. When some vectors are close to the centre,
47 this means that some information is carried on other axes, and that any interpretation might be hazardous.
48 The correlation circle is then used to identify the potential proximity with the 6 vectors and to assess their
49 potential correlations. Should a vector representing the variable "BT consumption" be close to the
50 correlation circle and point to a similar direction compared to any of the other 5 vectors representing
51 health indicators, this would indicate a positive correlation between the two variables.
52
53
54
55
56
57
58
59
60

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: $y = \mathbf{ax} + \mathbf{b}$ (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)

1
2
3 were represented with a large angle (close to orthogonal) compared to the BT vector, confirming poor
4 statistical relationships between BT and these 4 Health indicators. Of particular interest was the
5 interpretation of the "infectious disease" vector, which seemed to be close to the BT vector in a two
6 dimensions projection , but was actually represented by a large angle in the third dimension. The
7 infectious disease vector was also closer to the centre of the correlation circle, confirming the poor
8 meaningful correlations and potentially hazardous interpretations. Consequently, among the five health
9 indicators selected, only the diabetes parameter was correlated with BT consumption and can be
10 submitted to discussion and interpretation. No valid interpretations can be derived from the other four
11 health indicators using this dataset.

12
13
14
15
16
17 Then linear correlation model with the format $y = ax + b$ is represented by the following formula and is
18 presented in Figure 3:
19

$$\text{Diabetes prevalence} = - 0.0171183 * \text{BT consumption} + 6173.64$$

20
21
22 The y-coordinate of the point at which the regression line intersects the y-axis (intercept) can be
23 considered to correspond to the average prevalence of diabetes in a country in which BT consumption is
24 be unknown (6,173 cases per 100,000 inhabitants). Based on 42 countries, the p value of the Fisher-
25 Snedecor test was 0.003, which is highly significant, confirming the relevance of the linear model. The
26 coefficient r^2 was equal to 0.199. Student's t-test confirmed that the slope coefficient (0.0171183) was
27 significantly different from 0 ($p=0.003$) with a 5% confidence interval between [-0.028; -0.006]. The
28 negativity of coefficient "a" indicates that diabetes prevalence decreases as BT consumption increases,
29 confirming a negative correlation.
30
31
32
33
34
35
36

37 Discussion

38 39 40 *Limitations*

41
42 This study establishes an inverse linear statistical relationship between high BT consumption and diabetes
43 prevalence in the world, and confirms the findings of the European ecological study establishing a similar
44 relationship³². As in any database analysis, the very first limitation of this study is related to the quality of
45 the data. WHO prevalence data were obtained from the WHS, which constitutes a convenient and official
46 source of key morbidity indicators around the world. The general design of the WHS is based on
47 population sampling organized in the 192 Member States of the United Nations using face-to-face or
48 telephone interviews. As the survey questionnaire offers a menu of choices of modules for each country,
49 and lets the country select the survey approach (Household face-to-face survey, Computer-Assisted
50 Telephone Interview or Computer-Assisted Personal Interview), the quality of data collection can be
51 expected to be heterogeneous around the world.
52
53
54
55
56
57
58
59
60

1
2
3 Furthermore, some of the selected health indicators represent a group of diseases, such as infectious
4 diseases (tuberculosis and HIV) and cancer. The heterogeneity of these indicators can make it difficult to
5 establish any potential statistical relationships. Although more homogeneous, health indicators such as
6 diabetes depend on diagnostic criteria, which can vary from country to country. On the other hand, any
7 fixed survey design with fixed criteria would not be appropriate everywhere, for example in countries
8 with low telephone network coverage when planning telephone interviews. Other approaches to estimate
9 prevalence of diabetes in the world have been studies using literature and data extrapolations³³,
10 confirming the growing burden of diabetes. Another important concern is the interpretation of the
11 established statistical relationship between BT consumption and diabetes prevalence. Using a systematic
12 data mining approach, we tested the potential statistical relationship between BT consumption and 5
13 health indicators, without any *a priori* assumptions in relation to any of these health indicators. We
14 observed that, among the 5 health indicators, only the "prevalence of diabetes" indicator appeared to have
15 a strong statistical relationship with BT consumption. This ecological approach considers the population
16 as the unit of analysis rather than an individual, which is considered to be inferior to case-control studies
17 in the context of evidence-based medicine. In an ecological study, no information is available about the
18 individual members of the populations compared, whereas in a case-control study, information is reported
19 for each individual. A classical criticism of this approach is the "ecological fallacy", corresponding to a
20 logical fallacy in interpretation of the observed correlations at the population level, assuming that they
21 can be applied at the individual level. It is well known that statistics that accurately describe group
22 characteristics do not necessarily apply to individuals within that group. Our study on black tea does not
23 comprise any potential logical fallacy, as it was not used as the basis for any individual assumptions.
24 However, when interesting and strong associations are observed, the results of ecological studies have
25 provided numerous assumptions that have been subsequently confirmed by experimental studies. One of
26 the best known studies was that published by Keys in 1980³⁴ concerning the relationship with dietary
27 habits and coronary heart disease in 7 countries. The results of what later became known as the "Seven
28 Countries Study" appeared to show that serum cholesterol was strongly related to coronary heart disease
29 mortality at both the population and individual levels, leading to US government dietetic guidelines.
30 Other ecological studies have significantly contributed to scientific knowledge and public health
31 interventions, such as the relationship between lung cancer and tobacco, which has been confirmed by
32 numerous studies³⁵. For these reasons, ecological studies can be very useful for international
33 comparisons, while case-control studies are exclusively based on local information. Furthermore, when
34 strong correlations have been established, the results of ecological studies can suggest further evidence-
35 based studies, investigating the relevance and mechanism of the statistical relationship.
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

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 Europe³². 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 2030³³ 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 diseases³⁶. Furthermore, diabetic retinopathy is a common and specific microvascular complication of diabetes, and remains the leading cause of preventable blindness in working-aged people³⁷. With one of the highest prevalences of all human diseases, diabetes is now a global epidemic with devastating health, social and economic consequences³⁸. 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 populations³⁹. 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 burden⁴⁰. 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 Singapore⁴¹. 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 mixed⁴⁴ 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 day⁴⁹. This was confirmed by two meta-analyses published by Huxley *et al*⁴³ and Jing *et al*⁵⁰. Flavonoids are believed to support normal

1
2
3 glucose metabolism via anti-inflammatory effects and increased insulin activity^{51,52}. Various studies,
4 especially in Asian populations, confirm that flavonoids present in green tea could reduce fat absorption
5 in the gut, may promote fat oxidation in tissues and may increase energy expenditure⁵³. An observational
6 study of 4,300 Dutch adults found that flavonoid intake was highest in women who gained the least
7 weight over a 14-year period⁵⁴. Furthermore, as physical activity with or without diet contributes to a
8 healthier lifestyle, this important factor must be considered when comparing health indicators between
9 industrialized and emerging countries. Given rapid population growth, increased urbanization, and
10 adverse lifestyle changes, the obesity/type 2 diabetes epidemic in resource-poor nations was predicted in
11 the 1990s and has now been fully confirmed⁵⁵, underlying the importance of a better understanding of
12 predictive and potentially protective factors.
13
14
15
16
17

18 ***Correlation and causality***

19
20 The number of factors contributing to the growth of diabetes and obesity in the world confirms that
21 "correlation does not imply causality", and that a significant linear correlation between BT consumption
22 and diabetes prevalence does not imply that low BT consumption could *cause* diabetes. If one factor is
23 established as causing another, then the two factors are most certainly correlated. However, the opposite
24 cannot be concluded. Thus, a correlation can only indicate a potential direct or indirect possible cause,
25 which then needs to be further investigated. This paradigm and the connotations of causality may be the
26 most important considerations affecting biostatistics, not only in ecological studies but also in major
27 epidemiological study designs⁵⁶. A well known example of epidemiological cause-and-effect
28 misinterpretations is the correlation that was established between hormone replacement therapy and a
29 lower incidence of coronary heart disease. This association has been more recently explained by the fact
30 that women taking hormone replacement therapy were more likely to come from higher socio-economic
31 levels, which could explain the lower incidence of coronary heart disease⁵⁷. Establishing causality is one
32 of the most difficult challenges in public health. For instance, in clinical research, randomised controlled
33 clinical trials are performed to establish potential significant differences between two groups. However,
34 establishing a difference is not a demonstration of causality. Another example is case-control studies,
35 which compare individuals with a specific disease ("cases") with a group of individuals without the
36 disease ("controls"). An association between the hypothesized exposure and the disease studied would be
37 reflected by a higher proportion in exposed cases, but this cannot constitute a real demonstration of
38 causality. A potential causality can only be established with the convergence of interdisciplinary scientific
39 evidence (biological, physiological, epidemiological, etc.) and reasonable explanations based on
40 longitudinal studies. In any case, ecological research can address important issues that cannot be easily
41 addressed by other study designs. Ecological studies are frequently used when alternative study designs
42 are not possible (eg, randomised control trials), such as when investigating the effect of geographical
43 factors on disease incidence. Our research, like all ecological studies and most other epidemiological
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 approaches,presents a number of limitations because factors other than dietary habits may be the most
4 important determinants of variations in diabetes prevalence across communities. For example, it is
5 possible that other unmeasured confounding factors (eg, genetic differences) may explain some of the
6 observed regional variations. Due to the large number of potential determinants of diabetes prevalence,
7 including patient-, physician-, hospital-, and community-related variables, it is difficult to identify with
8 certainty all of the causes of the regional variations of diabetes prevalence, and additional follow-up
9 studies should be considered to confirm the hypotheses generated by this type of study. Despite the fact
10 that a number of biological, physiological and epidemiological field studies have provided evidence
11 linking BT consumption and glucose metabolism^{16, 17, 22, 42-44, 51, 52}, a large-scale randomised controlled
12 trial of tea consumption and diabetes risk would be useful to confirm these findings.
13
14
15
16
17

18 ***Data mining and data dredging***

19
20 Beyond the causality issue, a frequent criticism of using data mining was based on the confusion between
21 *data mining* and *data dredging* techniques. While a data mining approach is based on searching for
22 combinations of variables that might show potential correlations, data-dredging (also called "data
23 fishing") can generate misleading results⁵⁸. When a number of hypotheses are tested, it is expected that
24 some will falsely appear to be statistically significant, since every database can contain potential random
25 correlations. A robust data mining approach must therefore always be based on a clear research strategy
26 and a limited number of relevant meaningful assumptions. In our assessment, we used a systematic data
27 mining approach to test potential correlations between 6 selected variables (BT consumption and 5 key
28 health indicators). PCA was used to describe and structure the dataset before testing any correlations. In
29 our study, only one linear correlation model was constructed between BT consumption and diabetes
30 prevalence, based on the most relevant association suggested by the PCA. This consistent approach is
31 quite different from screening numerous cross-regression analyses between all variables of one particular
32 dataset. The data mining approach can be considered to be a "radar tracking system", allowing detection,
33 tracking and classification of potential "targets" in the framework of a particular environment. This is
34 particularly useful when exploring complex databases, as data mining can identify original statistical
35 evidence, which would never be discovered by means of classical statistical techniques. As an example,
36 the significant progress in genomics would not have been possible without the use of data mining
37 techniques. Despite the data collection homogeneity issue inherent to large cross-country comparisons,
38 we believe that this multidimensional approach can provide valuable additional scientific information,
39 completing published biological, physiological and epidemiological studies conducted on tea
40 consumption, diabetes and obesity. These results should support further causality research regarding the
41 health benefits of BT consumption on type 2 diabetes prevalence in the world.
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

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 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, Puangraphant 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, Chakravorty 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. Pełal A, Drózdź P, Biesaga M, et al. Evaluation of the antioxidant properties of fruit and flavoured black teas. *Eur J Nutr*. 2011;Mar 1.

- 1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
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. Psaltopoulou 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.

- 1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
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).

37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

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

1
2
3
4
5
6
7 **Relationships Between Black Tea Consumption and Key Health Indicators in**
8 **the World: an Ecological Study**

Formatted: Font: 14 pt, Bold

9
10 **~~Data Mining Approach to Assess Statistical Relationships Between Black Tea~~**
11 **~~Consumption and Key Health Indicators in the World~~**
12
13

14
15 Ariel Beresniak, MD, MPH, PhD, Data Mining International, Geneva, Switzerland
16 Gerard Duru, PhD, Data Mining International; Geneva, Switzerland
17 Genevieve Berger, MD, PhD, Unilever, London, UK
18 Dominique Bremond-Gignac, MD, PhD, Amiens University Hospital, Amiens, INSERM UMRS 968,
19 Paris VI University, France
20

21 **Corresponding author:** The corresponding author has the right to grant on behalf of all authors and does
22 grant on behalf of all authors, an exclusive licence on a worldwide basis to the BMJ Publishing Group Ltd
23 to permit this article to be published in BMJ Open and any other BMJPGJL products and sublicences such
24 use and to exploit all subsidiary rights, as set out in our licence.
25

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

35 **Funding statements:** This research received no specific grant from any funding agency in the public,
36 commercial or not-for-profit sectors.
37

38 **Competing interest statement:** All authors have completed the Unified Competing Interest form: no
39 support from any organisation for the submitted work, no financial relationships with any organisations
40 that might have an interest in the submitted work in the previous three years. Coauthor G.B is employed
41 at Unilever PLC as Chief R&D Officer and has provided access to the Euromonitor International global
42 tea consumption data without any financial agreement or any grant to support this study, which has been
43 carried out in total independence. There are no other relationships or activities that could appear to have
44 influenced the submitted work.
45

46 **Acknowledgements:** The authors thank Professor Paul Zimmet, Director Emeritus, Baker IDI Heart and
47 Diabetes Institute, for reviewing the manuscript and making comments.
48

49 **Data sharing statements:** This study used epidemiological data from WHO World Health Surveys.
50

51 **Individual contributions:** The work presented here was carried out in collaboration between all authors.
52 A.B. conceived the study aims and design. G.B. contributed to the data collection. A.B. and G.D
53 performed the analysis. A.B, G.D and D.B. interpreted the results. A.B drafted the manuscript. All
54 authors have contributed to, seen and approved the manuscript.
55
56
57
58
59
60

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60**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 randomized 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 ecological study carried out in European 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* experiments¹⁻⁹. 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 disease¹⁰, 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 actions¹¹ 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 effects¹³. Moreover, dietary consumption of tea catechins would have beneficial effects on blood pressure and lipid parameters¹⁴. 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 cancer¹⁵. 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 beverages¹⁸⁻²¹. The relatively recent interest in BT may be explained by the fact that BT is historically the type of tea

4

1
2
3
4
5
6
7 most widely consumed in Western countries, probably due to its good storage properties, promoting
8 active trade with tea-producing countries in Asia. Although there has recently been a renewed interest in
9 green tea in industrialized countries, BT represents over ninety percent of all tea sold in the West. Despite
10 the number of publications investigating the effects of tea components and green tea in particular, large-
11 scale clinical dose-effect studies are still lacking and it is difficult to interpret the clinical significance of
12 results derived from some biological studies. Considerably fewer studies have been conducted
13 specifically on BT, mostly investigating its antioxidant properties^{22, 23}, and cardiovascular effects^{24, 25}. It
14 could therefore be interesting to investigate some key dietary habits in relation with lifestyle and health
15 effects at a global level, in view of the perceived positive health effects of black tea (BT), which have
16 been described for centuries^{26, 27}. Because of the complexity of implementing international prospective
17 studies and the difficulty of conducting meta-analyses on a large number of heterogeneous local studies,
18 potential correlations between BT consumption and epidemiological data around the world could be
19 investigated by using advanced exploratory statistical methods. The objective of this original research
20 was to investigate potential statistical relationships between BT consumption and the following five key
21 health indicators: respiratory diseases, infectious diseases, cancer, cardiovascular diseases and diabetes.
22 An almost 6-fold increase in the number of people with diabetes has been observed over the last few
23 decades. The International Diabetes Federation (IDF) reports that the number of people with diabetes will
24 escalate from 285 million to 438 million between 2010 and 2030¹ and the number of persons with IGT
25 will increase from 344 to 472 million. By 2030, there will be over 900 million people worldwide with
26 diabetes or at high risk of diabetes. Diabetes confers about a two-fold excess risk for a wide range of
27 vaseular diseases². Furthermore, diabetic retinopathy is a common and specific microvascular
28 complication of diabetes, and remains the leading cause of preventable blindness in working-aged people
29 ³. With one of the highest prevalences of all human diseases, diabetes is now a global epidemic with
30 devastating health, social and economic consequences⁴. In certain ethnic groups, such as Asian
31 populations, diabetes develops at a younger age than in Caucasian populations. Several distinctive
32 features are apparent in the pathogenic factors for diabetes and their thresholds in Asian populations⁵. In
33 conjunction with genetic susceptibility, type 2 diabetes is brought on by environmental and behavioural
34 factors such as a sedentary lifestyle, overly rich nutrition and obesity and results in a huge economic
35 burden⁶. It could therefore be interesting to investigate some key dietary habits in relation with lifestyle
36 and health effects at a global level. For example, the positive health effects of black tea (BT) have been
37 observed for centuries⁷⁻⁸.
38 Considering the complexity of implementing international prospective studies and the difficulty of
39 conducting meta-analyses on a large number of heterogeneous local studies, potential correlations
40 between BT consumption and epidemiological data around the world could be investigated by deploying
41 a data-mining approach using advanced exploratory statistical methods.
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

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 ²⁸⁹. 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 ²⁹⁰. 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).

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

Field Code Changed

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."

~~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 assessment³⁴⁺⁵. 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 F2³⁵⁺⁶. 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.

Formatted: Line spacing: 1.5 lines

Field Code Changed

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)

1
2
3
4
5
6
7 were represented with a large angle (close to orthogonal) compared to the BT vector, confirming poor
8 statistical relationships between BT and these 4 Health indicators. Of particular interest was the
9 interpretation of the "infectious disease" vector, which seemed to be close to the BT vector in a two
10 dimensions projection, but was actually represented by a large angle in the third dimension. The
11 infectious disease vector was also closer to the centre of the correlation circle, confirming the poor
12 meaningful correlations and potentially hazardous interpretations. Consequently, among the five health
13 indicators selected, only the diabetes parameter was correlated with BT consumption and can be
14 submitted to discussion and interpretation. No valid interpretations can be derived from the other four
15 health indicators using this dataset.

16
17
18
19 Then linear correlation model with the format $y = ax + b$ is represented by the following formula and is
20 presented in Figure 3:

$$\text{Diabetes prevalence} = -0.0171183 * \text{BT consumption} + 6173.64$$

21
22
23
24 The y-coordinate of the point at which the regression line intersects the y-axis (intercept) can be
25 considered to correspond to the average prevalence of diabetes in a country in which BT consumption is
26 be unknown (6,173 cases per 100,000 inhabitants). Based on 42 countries, the p value of the Fisher-
27 Snedecor test was 0.003, which is highly significant, confirming the relevance of the linear model. The
28 coefficient r^2 was equal to 0.199. Student's t-test confirmed that the slope coefficient (0.0171183) was
29 significantly different from 0 ($p=0.003$) with a 5% confidence interval between [-0.028; -0.006]. The
30 negativity of coefficient "a" indicates that diabetes prevalence decreases as BT consumption increases,
31 confirming a negative correlation.

32
33
34
35 The database was composed of 300 fields representing 6 variables (5 health indicators and BT
36 consumption) in 50 countries. Using normative PCA on this multidimensional table, the overall "quality"
37 (percentage of original variance) of the final projection from 6 dimensions (6 variables) was 59% in 2
38 dimensions and 74% when projected in 3 dimensions. This confirms that the best representation of the
39 dataset should be in 3 dimensions, which can be described by axes entitled F1, F2 and F3. The "BT
40 consumption" variable provided a high contribution to the construction on axis F3 (81%). The angle of
41 the vector "BT consumption" with axis F3 was only 22°, confirming the very high contribution of this
42 variable on axis F3. Forty two of the 50 countries were related to this axis F3. The 8 countries not
43 contributing to F3 were Brazil, China, Venezuela, Morocco, Colombia, Vietnam, Philippines and Israel,
44 suggesting the absence of any correlation between BT consumption and health indicators in these
45 particular countries.

46
47
48
49 The "correlation circle" (Figure 3) shows that the "BT consumption" vector was strictly opposite the
50 "Diabetes prevalence" vector, establishing a strong statistical negative correlation. Vectors concerning the
51 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:

$$\text{Diabetes prevalence} = a * \text{BT consumption} + b$$

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).

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

$$\text{Diabetes prevalence} = -0.0171183 * \text{BT consumption} + 6173.64$$

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. 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, Bold, Italic

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

Formatted: Font: Bold, Italic

Formatted: Font: 11 pt

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

Formatted: Font color: Text 1

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 extrapolations³⁶⁺ confirming the growing burden of diabetes.

Field Code Changed

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 1980³⁷ 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 studies³⁸. 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: 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

Formatted: Font: 11 pt

Formatted: Superscript

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 beverages²²⁻²⁵. 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. 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 2030³⁶ 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 diseases⁴⁰. Furthermore, diabetic retinopathy is a common and specific microvascular complication of diabetes, and remains the leading cause of preventable blindness in working-aged people⁴¹. With one of the highest prevalences of all human diseases, diabetes is now a global epidemic with devastating health, social and economic consequences⁴². 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 populations⁴³. 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 burden⁴⁴. 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 Singapore⁴⁵³⁶.

Although many laboratory studies have observed physiological effects of BT on glucose metabolism^{17,18,46,4731,32,37,38}, the underlying mechanisms remain unclear. The results of human intervention studies are mixed⁴⁸³⁹ and the role of caffeine has been suggested but not clearly established^{47,4938,40}. Neyestani *et al*⁴⁶³⁷ 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*¹⁷³¹ 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*⁵⁰⁴⁴ 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^{51,5243,43}.

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 day⁵³⁴⁴. This was confirmed by two meta-analyses published by Huxley *et al*⁴⁷³⁸ and Jing *et al*⁵⁴⁴⁵. Flavonoids are believed to support normal glucose metabolism via anti-inflammatory effects and increased insulin activity^{55,5646,47}. 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 expenditure⁵⁷⁴⁸. 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 period⁵⁸⁴⁹. 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

1
2
3
4
5
6
7 contributes to a healthier lifestyle, this important factor must be considered when comparing health
8 indicators between industrialized and emerging countries. Given rapid population growth, increased
9 urbanization, and adverse lifestyle changes, the obesity/type 2 diabetes epidemic in resource-poor nations
10 was predicted in the 1990s and has now been fully confirmed^{59,60}, underlying the importance of a better
11 understanding of predictive and potentially protective factors.

Field Code Changed

12 Correlation and causality

Formatted: Font: Bold, Italic

13
14 The number of factors contributing to the growth of diabetes and obesity in the world confirms that
15 "correlation does not imply causality", and that a significant linear correlation between BT consumption
16 and diabetes prevalence does not imply that low BT consumption could cause diabetes. If one factor is
17 established as causing another, then the two factors are most certainly correlated. However, the opposite
18 cannot be concluded. Thus, a correlation can only indicate a potential direct or indirect possible cause,
19 which then needs to be further investigated. This paradigm and the connotations of causality may be the
20 most important considerations affecting biostatistics, not only in ecological studies but also in major
21 epidemiological study designs.⁶⁰ A well known example of epidemiological cause-and-effect
22 misinterpretations is the correlation that was established between hormone replacement therapy and a
23 lower incidence of coronary heart disease. This association has been more recently explained by the fact
24 that women taking hormone replacement therapy were more likely to come from higher socio-economic
25 levels, which could explain the lower incidence of coronary heart disease.⁶¹ Establishing causality is one
26 of the most difficult challenges in public health. For instance, in clinical research, randomised controlled
27 clinical trials are performed to establish potential significant differences between two groups. However,
28 establishing a difference is not a demonstration of causality. Another example is case-control studies,
29 which compare individuals with a specific disease ("cases") with a group of individuals without the
30 disease ("controls"). An association between the hypothesized exposure and the disease studied would be
31 reflected by a higher proportion in exposed cases, but this cannot constitute a real demonstration of
32 causality. A potential causality can only be established with the convergence of interdisciplinary scientific
33 evidence (biological, physiological, epidemiological, etc.) and reasonable explanations based on
34 longitudinal studies. In any case, ecological research can address important issues that cannot be easily
35 addressed by other study designs. Ecological studies are frequently used when alternative study designs
36 are not possible (eg, randomised control trials), such as when investigating the effect of geographical
37 factors on disease incidence. Our research, like all ecological studies and most other epidemiological
38 approaches, presents a number of limitations because factors other than dietary habits may be the most
39 important determinants of variations in diabetes prevalence across communities. For example, it is
40 possible that other unmeasured confounding factors (eg, genetic differences) may explain some of the
41 observed regional variations. Due to the large number of potential determinants of diabetes prevalence,
42 including patient-, physician-, hospital-, and community-related variables, it is difficult to identify with
43
44
45
46
47
48
49
50
51
52

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.

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 designs⁵⁴. 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 disease⁵². 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

Formatted: Superscript

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.

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.

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

6. de Meija E, Ramirez-Mares M, Puangpraphant S. Bioactive components of tea: cancer, inflammation and behavior. *Brain Behav Immun*. 2009;23(6):721-731. Formatted ... [1]
7. Béliveau R, Gingras D. Green tea: prevention and treatment of cancer by nutraceuticals. *Lancet*. 2004;364(9439):1021-1022. Formatted ... [2]
8. Walsh G. Tea and heart disease. *Lancet*. 1997;349(9053):735. Formatted ... [3]
9. Ras R, Zock P, Draijer R. Tea Consumption Enhances Endothelial-Dependent Vasodilation; a Meta-Analysis. *PLoS ONE* 2011;6(3):e16974. Formatted ... [4]
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. Formatted ... [5]
11. Marathe SA, Datey AA, Chakravorty D. Herbal Cocktail as Anti-infective: Promising Therapeutic for the Treatment of Viral Diseases. *Recent Pat Antiinfect Drug Discov*. 2012;7(2):123-132. Formatted ... [6]
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. Formatted ... [7]
13. Islam MA. Cardiovascular effects of green tea catechins: progress and promise. *Recent Pat Cardiovasc Drug Discov*. 2012;7(2):88-99. Formatted ... [8]
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. Formatted ... [9]
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. Formatted ... [10]
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. Formatted ... [11]
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. Formatted ... [12]
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. Formatted ... [13]
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. Formatted ... [14]
20. Yoshioka K, Kogure A, Yoshida T, Yoshikawa T. Coffee consumption and risk of type 2 diabetes mellitus. *Lancet*. 2003;361(9358):703. Formatted ... [15]
21. Reunanen A, Heliövaara M, Aho K. Coffee consumption and risk of type 2 diabetes mellitus. *Lancet*. 2003;361(9358):702-703. Formatted ... [16]
22. Pękal A, Drózdź P, Biesaga M, Pyrzynska K. Evaluation of the antioxidant properties of fruit and flavoured black teas. *Eur J Nutr*. 2011;Mar 1. Formatted ... [17]
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. Formatted ... [18]
24. Baborun 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. Formatted ... [19]
25. Tokudome S, Nahomi I, Goto C, Tokudome Y, Moore M. Black tea and cardiovascular disease. *Int J Epidemiol*. 2005;34(2):482-483. Formatted ... [20]
26. Baborun 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. Formatted ... [21]
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;93(3):506-515. Formatted ... [22]
28. Euromonitor. Hot Drinks: trade sources. 2010;www.euromonitor.com. Formatted ... [23]
29. WHO. Global Health Survey. 2009;http://www.who.int/healthinfo/survey/en/. Formatted ... [24]
30. Naqa I, Deasy J, Mu Y, et al. Datamining approaches for modeling tumor control probability. *Acta Oncol*. 2010;49(8):1363-1373. Formatted ... [25]

- 1
2
3
4
5
6
7 **31.** Zhang F, Chen J. Data mining methods in omics-based biomarker discovery. *Methods Mol Biol.* 2011(719):511-526. Formatted ... [26]
- 8 **32.** Wei CK SS, Yang MC. Application of Data Mining on the Development of a Disease
9 Distribution Map of Screened Community Residents of Taipei County in Taiwan. *J Med Syst.*
10 2011(Feb 25). Formatted ... [27]
- 11 **33.** Harpaz R, Haerian K, Chase H, Friedman C. Statistical Mining of Potential Drug Interaction
12 Adverse Effects in FDA's Spontaneous Reporting System. *AMIA Annu Symp Proc.* 2010(Nov
13 13):281-285. Formatted ... [28]
- 14 **34.** Briand S, Beresniak A, Nguyen T, et al. Assessment of yellow fever epidemic risk: an original
15 multi-criteria modeling approach. *PLoS Negl Trop Dis.* 2009;3(7):e483. Formatted ... [29]
- 16 **35.** Everitt B, Dunn G. Applied Multivariate data analysis. *Lavoisier Publisher.* 2001, 2d ed.:320p Formatted ... [30]
- 17 **36.** Shaw JE, Sicree RA, Zimmet PZ. Global estimates of the prevalence of diabetes for 2010 and
18 2030. *Diabetes Research and Clinical Practice.* 2010(87):4-14. Formatted: Indent: Left: 0", First line: 0" Formatted ... [31]
- 19 **37.** Keys A. Seven Countries: A Multivariate Analysis of Death and Coronary Heart Disease.
20 *Harvard University Press.* 1980. Formatted ... [32]
- 21 **38.** Didkowska J, Manczuk M, McNeill A, Powles J, Zatonski W. Lung cancer mortality at ages 35-
22 54 in the European Union: ecological study of evolving tobacco epidemics. *BMJ.*
23 2005;331(7510):189-191. Formatted ... [33]
- 24 **39.** Consortium. TI. Tea Consumption and Incidence of Type 2 Diabetes in Europe: The EPIC-
25 InterAct Case-Cohort Study. *PLoS One.* 2012(7):5. Formatted ... [34]
- 26 **40.** Emerging Risk Factors Collaboration, Sarwar N, Gao P, et al. Diabetes mellitus, fasting blood
27 glucose concentration, and risk of vascular disease: a collaborative meta-analysis of 102
28 prospective studies. *Lancet.* 2010;375(9733):2215-2222. Formatted ... [35]
- 29 **41.** Cheung N, Mitchell P, Wong TY. Diabetic retinopathy. *Lancet.* 2010;376(9735):124-136. Formatted ... [36]
- 30 **42.** Danaei G, Finucane MM, Lu Y, et al. National, regional, and global trends in fasting plasma
31 glucose and diabetes prevalence since 1980: systematic analysis of health examination surveys
32 and epidemiological studies with 370 country-years and 2·7 million participants. *Lancet.*
33 2011;10.1016/S0140-6736(11):60679-X. Formatted ... [37]
- 34 **43.** Ramachandran A MR, Snehalatha C. Diabetes in Asia. *Lancet.* 2010;375(9712):408-418. Formatted ... [38]
- 35 **44.** Zimmet P, Alberti KG, Shaw J. Global and societal implications of the diabetes epidemic. *Nature*
36 *Reviews.* 2001;414(6865):782-787. Formatted ... [39]
- 37 **45.** Ang YG, Wu XC, Toh MP, Chia KS, Heng BH. Progression Rate of newly diagnosed Impaired
38 Fasting Glycemia to Type 2 Diabetes Mellitus: a study using the National Healthcare Group
39 Diabetes Registry in Singapore. *J Diabetes.* 2011;Nov 7. Formatted ... [40]
- 40 **46.** Nevestani T, Shariatzade N, Kalayi A, et al. Regular daily intake of black tea improves oxidative
41 stress biomarkers and decreases serum C-reactive protein levels in type 2 diabetic patients. *Ann*
42 *Nutr Metab.* 2010;57(1):40-49. Formatted ... [41]
- 43 **47.** Huxley R, Lee C, Barzi F, et al. Coffee, decaffeinated coffee, and tea consumption in relation to
44 incident type 2 diabetes mellitus: a systematic review with meta-analysis. *Arch Intern Med.*
45 2009;169(22):2053-2063. Formatted ... [42]
- 46 **48.** Hayashino Y, Fukuhara S, Okamura T, Tanaka T, Ueshima H, Group. H-OR. High oolong tea
47 consumption predicts future risk of diabetes among Japanese male workers: a prospective cohort
48 study. *Diabet Med.* 2011(Jan 18). Formatted ... [43]
- 49 **49.** Goto A, Song Y, Chen B, Manson J, Buring J, Liu S. Coffee and caffeine consumption in relation
50 to sex hormone-binding globulin and risk of type 2 diabetes in postmenopausal women. *Diabetes.*
51 2011;60(1):269-275. Formatted ... [44]
- 52 **50.** Psaltopouloum T, Ilias I, Alevizaki M. The role of diet and lifestyle in primary, secondary, and
53 tertiary diabetes prevention: a review of meta-analyses. *Rev Diabet Stud.* . 2010;7(1):26-35. Formatted ... [45]
- 54 **51.** Martini L, Catania A, Ferreira S. Role of vitamins and minerals in prevention and management of
55 type 2 diabetes mellitus. *Nutr Rev.* 2010;68(6):341-354. Formatted ... [46]
- 56 **52.** Suksomboon N, Poolsup N, Sinprasert S. Effects of vitamin E supplementation on glycaemic
57 control in type 2 diabetes: systematic review of randomized controlled trials. *J Clin Pharm Ther.*
58 2011;36(1):53-63. 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. **Formatted: Font: Times New Roman, Bold**
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. **Formatted: Font: Times New Roman**
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. **Formatted: Font: Times New Roman, Italic**
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. **Formatted: Font: Times New Roman**
57. MS. W-P. Green tea catechins, caffeine and body-weight regulation. *Physiol Behav*. 2010;100(1):42-46. **Formatted: Font: Times New Roman, Bold**
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. **Formatted: Font: Times New Roman, Italic**
59. Nour N. Obesity in resource-poor nations. *Rev Obstet Gynecol*. 2010;3(4):180-184. **Formatted: Font: Times New Roman**
60. Ortega Calvo M, Román Torres P, Lapetra Peralta J. Epistemology as health research propedeutics. *Gac Sanit*. 2011. **Formatted: Font: Times New Roman, Bold**
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. **Formatted: Font: Times New Roman, Italic**
62. Lord S, GebSKI V, Keech A. Multiple analyses in clinical trials: sound science or data dredging? *Med J Aust*. 2004;181(18). **Formatted: Font: Times New Roman**
1. 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. **Formatted: Font: Times New Roman, Bold**
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. **Formatted: Font: Times New Roman, Italic**
3. Cheung N, Mitchell P, Wong TY. Diabetic retinopathy. *Lancet*. 2010;376(9735):124-136. **Formatted: Font: Times New Roman**
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. **Formatted: Font: Times New Roman, Bold**
5. Ramachandran A MR, Snehalatha C. Diabetes in Asia. *Lancet*. 2010;375(9712):408-418. **Formatted: Font: Times New Roman**
6. Zimmet P, Alberti KG, Shaw J. Global and societal implications of the diabetes epidemic. *Nature Reviews*. 2001;414(6865):782-787. **Formatted: Font: Times New Roman, Italic**
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. **Formatted: Font: Times New Roman, Bold**
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. **Formatted: Font: Times New Roman**
9. Euromonitor. Hot Drinks: trade sources. 2010;www.euromonitor.com. **Formatted: Font: Times New Roman, Italic**
10. WHO. Global Health Survey. 2009;http://www.who.int/healthinfo/survey/en/. **Formatted: Font: Times New Roman**
11. Naqa I, Deasy J, Mu Y, et al. Data mining approaches for modeling tumor control probability. *Acta Oncol*. 2010;49(8):1363-1373. **Formatted: Font: Times New Roman, Bold**
12. Zhang F, Chen J. Data mining methods in omics-based biomarker discovery. *Methods Mol Biol*. 2011(719):511-526. **Formatted: Font: Times New Roman, Italic**
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). **Formatted: Font: Times New Roman**
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: 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
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, Zoek P, Draijer R. Tea Consumption Enhances Endothelial-Dependent Vasodilation: a Meta-Analysis. *PLoS ONE* 2011;6(3):e16974.
26. Pękal A, Drózd 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.

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

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. Psaltopoulou 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. Nicoelle 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).

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

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.)

<u>Hungary</u>	<u>2505</u>	<u>62</u>	<u>2204</u>	<u>4685</u>	<u>5927</u>	<u>11270</u>
<u>Ukraine</u>	<u>2552</u>	<u>857</u>	<u>2245</u>	<u>4630</u>	<u>4612</u>	<u>32290</u>
<u>Turkey</u>	<u>2931</u>	<u>48</u>	<u>1271</u>	<u>1579</u>	<u>3326</u>	<u>166310</u>
<u>Egypt</u>	<u>3121</u>	<u>40</u>	<u>615</u>	<u>1316</u>	<u>3979</u>	<u>95910</u>
<u>Saudi Arabia</u>	<u>3221</u>	<u>54</u>	<u>353</u>	<u>914</u>	<u>4257</u>	<u>57020</u>

For peer review only

8/15/2012 6:51:00 PM

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

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

For peer review only

8/15/2012 6:51:00 PM

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

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

For peer review only

8/15/2012 6:51:00 PM

- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8
- 9
- 10
- 11
- 12
- 13
- 14
- 15
- 16
- 17
- 18
- 19
- 20
- 21
- 22
- 23
- 24
- 25
- 26
- 27
- 28
- 29
- 30
- 31
- 32
- 33
- 34
- 35
- 36
- 37
- 38
- 39
- 40
- 41
- 42
- 43
- 44
- 45
- 46
- 47
- 48
- 49
- 50
- 51
- 52
- 53
- 54
- 55
- 56
- 57
- 58
- 59
- 60

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

For peer review only

8/15/2012 6:51:00 PM

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

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

For peer review only

8/15/2012 6:51:00 PM

- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8
- 9
- 10
- 11
- 12
- 13
- 14
- 15
- 16
- 17
- 18
- 19
- 20
- 21
- 22
- 23
- 24
- 25
- 26
- 27
- 28
- 29
- 30
- 31
- 32
- 33
- 34
- 35
- 36
- 37
- 38
- 39
- 40
- 41
- 42
- 43
- 44
- 45
- 46
- 47
- 48
- 49
- 50
- 51
- 52
- 53
- 54
- 55
- 56
- 57
- 58
- 59
- 60

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

For peer review only

8/15/2012 6:51:00 PM

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

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

For peer review only

8/15/2012 6:51:00 PM

- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8
- 9
- 10
- 11
- 12
- 13
- 14
- 15
- 16
- 17
- 18
- 19
- 20
- 21
- 22
- 23
- 24
- 25
- 26
- 27
- 28
- 29
- 30
- 31
- 32
- 33
- 34
- 35
- 36
- 37
- 38
- 39
- 40
- 41
- 42
- 43
- 44
- 45
- 46
- 47
- 48
- 49
- 50
- 51
- 52
- 53
- 54
- 55
- 56
- 57
- 58
- 59
- 60

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

For peer review only

8/15/2012 6:51:00 PM

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

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

For peer review only

8/15/2012 6:51:00 PM

- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8
- 9
- 10
- 11
- 12
- 13
- 14
- 15
- 16
- 17
- 18
- 19
- 20
- 21
- 22
- 23
- 24
- 25
- 26
- 27
- 28
- 29
- 30
- 31
- 32
- 33
- 34
- 35
- 36
- 37
- 38
- 39
- 40
- 41
- 42
- 43
- 44
- 45
- 46
- 47
- 48
- 49
- 50
- 51
- 52
- 53
- 54
- 55
- 56
- 57
- 58
- 59
- 60

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

For peer review only

8/15/2012 6:51:00 PM

- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8
- 9
- 10
- 11
- 12
- 13
- 14
- 15
- 16
- 17
- 18
- 19
- 20
- 21
- 22
- 23
- 24
- 25
- 26
- 27
- 28
- 29
- 30
- 31
- 32
- 33
- 34
- 35
- 36
- 37
- 38
- 39
- 40
- 41
- 42
- 43
- 44
- 45
- 46
- 47
- 48
- 49
- 50
- 51
- 52
- 53
- 54
- 55
- 56
- 57
- 58
- 59
- 60

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

For peer review only

8/15/2012 6:51:00 PM

- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8
- 9
- 10
- 11
- 12
- 13
- 14
- 15
- 16
- 17
- 18
- 19
- 20
- 21
- 22
- 23
- 24
- 25
- 26
- 27
- 28
- 29
- 30
- 31
- 32
- 33
- 34
- 35
- 36
- 37
- 38
- 39
- 40
- 41
- 42
- 43
- 44
- 45
- 46
- 47
- 48
- 49
- 50
- 51
- 52
- 53
- 54
- 55
- 56
- 57
- 58
- 59
- 60

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

For peer review only

8/15/2012 6:51:00 PM

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

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

For peer review only

8/15/2012 6:51:00 PM

- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8
- 9
- 10
- 11
- 12
- 13
- 14
- 15
- 16
- 17
- 18
- 19
- 20
- 21
- 22
- 23
- 24
- 25
- 26
- 27
- 28
- 29
- 30
- 31
- 32
- 33
- 34
- 35
- 36
- 37
- 38
- 39
- 40
- 41
- 42
- 43
- 44
- 45
- 46
- 47
- 48
- 49
- 50
- 51
- 52
- 53
- 54
- 55
- 56
- 57
- 58
- 59
- 60

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

For peer review only

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

For peer review only

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

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

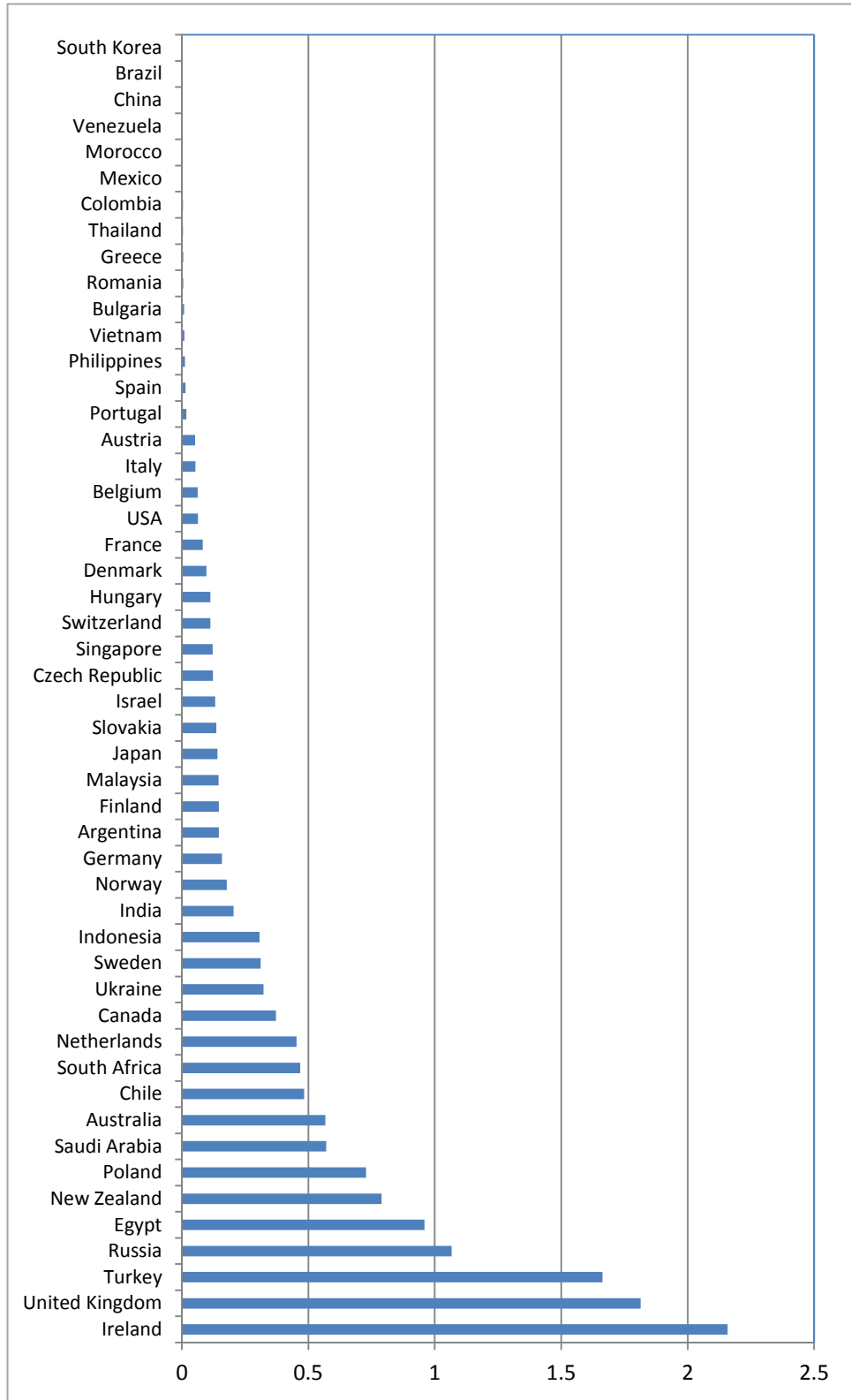
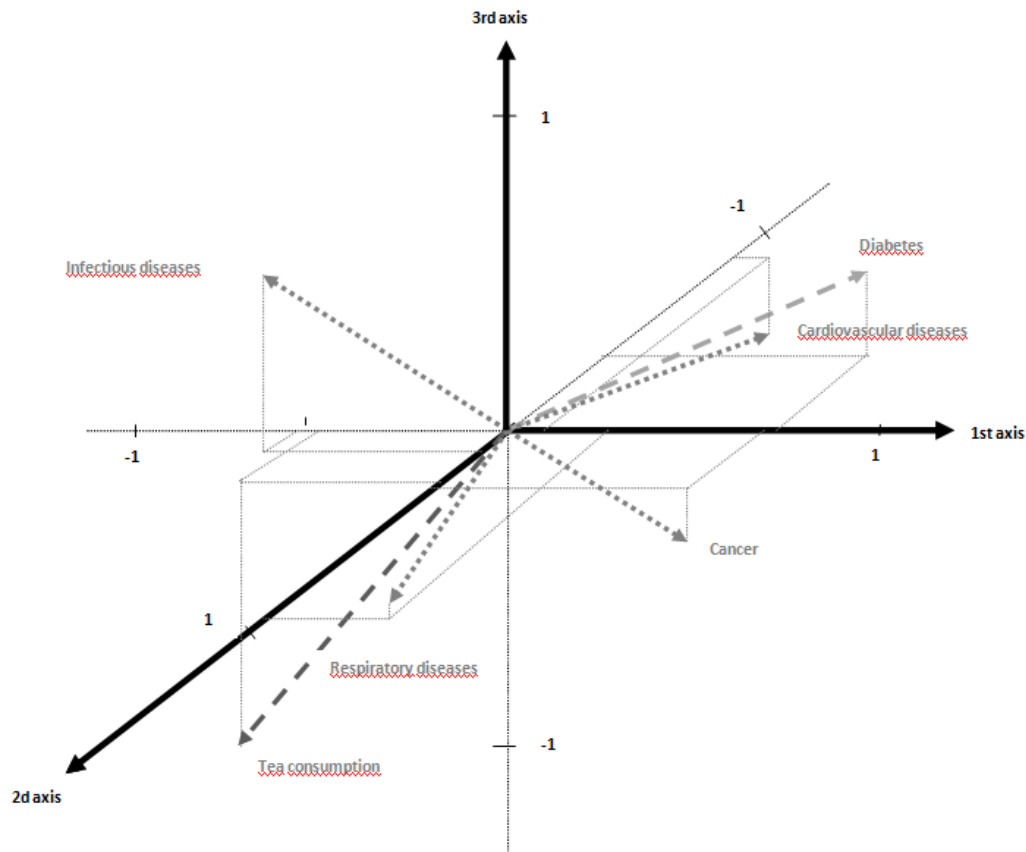


Figure 2: Three dimensional correlation circle of 5 health indicators and BT consumption*



*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)

