Research Paper

A study of the effect of water quality-related variables on some age-related diseases adjusted for other well-known risk factors: a multivariate multilevel study

Kapuruge Nishika Oshadini Ranathunga and M. R. Sooriyarachchi

ABSTRACT

Mortality rates of some diseases are affected by water quality. This research examines the roles of two factors related to water quality, namely the quality of drinking water termed 'water' and the quality of sanitation termed 'sanitation'. Two age-related diseases, cardiovascular disease and diabetes (CDD) and chronic respiratory conditions (CRC) are considered while adjusting for personal health issues, environmental and geographical factors. The dataset consists of worldwide mortality rates of adults for the mentioned diseases in 195 countries. These countries are clustered within continents geographically and literature shows the importance of considering the geographical effect of a continent. Furthermore, the two diseases were highly related to each other. Accordingly, the multivariate multilevel model was fitted to the dataset. The results indicated that when the usage of improved drinking water sources and sanitation facilities decreases, the chance of mortality from the two diseases increases. Furthermore, the difference in the risk of the diseases was statistically significant between the continents. It showed that North America and Europe had a lower risk of having CDD and CRC compared to Asia and Oceania. Therefore, the results revealed that the factors 'water' and 'sanitation' play important roles for this macro geographical variation of CDD and CRC.

Key words | cardiovascular disease and diabetes, chronic respiratory conditions, Markov chain Monte Carlo, multivariate multilevel model, probit regression

Kapuruge Nishika Oshadini Ranathunga
M. R. Sooriyarachchi (corresponding author)
Department of Statistics,
University of Colombo,
PO Box 1490,
Colombo 03,
Sri Lanka
E-mail: *roshinis@hotmail.com*

INTRODUCTION

Vulnerability to age-associated diseases is caused by aging. Some of these diseases may aggregate mortality among adults worldwide. Two such diseases are 'cardiovascular diseases and diabetes (CDD)' and 'chronic respiratory conditions (CRC)'. These are fatal diseases with increasing occurrence over time.

Cardiovascular disease is a type of disease which involves the heart, the blood vessels or both. Cardiovascular deaths have increased greatly in low and middle-income countries. It is accounting for 17 million deaths per year globally and is expected to grow to more than 23.6 million by 2030 (Go *et al.* 2014). Diabetes also has a considerable contribution to the worldwide mortality rates of adults in recent years and may occur when the insulin production in the body is insufficient, or the living cells in the body do not respond well to insulin, or both. It accounts for approximately 4 million deaths worldwide in 2012 according to WHO reports (Fuster & Kelly 2010).

Although CDD are two different diseases, they have a strong relationship (Ranathunga & Sooriyarachchi 2017). When considering the CRC, respiratory conditions are the most commonly managed problems in general practice. These affect the airways, including the lungs as well as the paths that transfer air from the mouth and nose into the

doi: 10.2166/washdev.2017.118

lungs. Asthma, lung cancer, occupational lung disease and chronic obstructive pulmonary disease are some of the most common diseases of these types.

Several studies have been carried out on the relationship between CDD and CRC with other health-related factors such as smoking, alcohol, obesity, blood sugar, cholesterol, hypertension, etc. The authors of this paper have used the data used here for such a study (Ranatunga & Sooriyarachchi 2017). A previous study (Ranatunga & Sooriyarachchi 2017) conducted by the same authors is a highly technical study of multivariate modeling and its application to the data explained before in this section and is suitable mainly for an audience familiar with advanced statistics. The current paper is an applied one addressing a different aspect to the same problem. This paper is more approriate for the medical community, including epidemiologists.

There is little literature showing that the factors, quality of drinking water and quality of sanitation also affect CDD and CRC, even causing death in the long-term (Briggs 2003). On the other hand, CDD and CRC are diseases with globally higher rates of mortality. The World Health Organization collects many details on these diseases and water and sanitation. However, there are no proper statistical and health information systems in many countries to study the effect of water-related factors on death due to CDD and CRC and hence the effort has not been made to ensure an end to this tragedy. This is the motivation to undertake this study involved in promoting facilities, mainly regarding water and sanitation, to control this vulnerable situation.

This study mainly aims to determine the effect of the factors 'water' and 'sanitation' on the world-wide mortality rates of CDD and CRC, while adjusting for the risk factors 'solid fuels', 'blood glucose', 'blood pressure', 'smoking', 'alcohol consumption' and 'obesity'.

METHODS

Dataset

The book 'World Health Statistics (2013)' published by the World Health Organization (WHO) in 2013 was used to extract the data. Worldwide mortality rates among adults aged 30–70 years were represented by the dataset. These observations were collected as frequencies per 100,000 population. Data were spread across two main levels. Level 1 consists of 195 countries while level 2 consists of six continents. The countries are believed to be similar within continents but vary across continents (http://education.nationalgeographic.com/encyclopedia/continent/). This establishes the need for multilevel analysis, as the data represent a hierarchical nature.

Statistical methods

Initially, bar charts and mosaic plots were used and it was followed by the Generalized Cochran Mantel–Haenszel (GCMH) univariate test to assess individual significance. After that, by considering continent correlations and strong relationships between the two diseases, multivariate multilevel modeling was implemented.

In order to conduct preliminary analysis, all variables were categorized according to their percentiles due to the lack of methods about handling continuous data in a hierarchical nature. Moreover, for advanced analysis, it was not possible to model continuous responses since these deviated from Normality and Multivariate Normality. Therefore, advanced analysis was also performed by using categorized response variables.

Graphical tools

For the graphical phase, mosaic plots were used (Friendly 1994). Mosaic plots are one of the most powerful visualization tools for multivariate categorical data. These are constructed based on the data in the contingency tables and it can be extended up to four-way contingency tables. The areas of the tiles of the plot are proportional to the cell frequencies of the contingency table. To identify the pattern of deviation from the independence, Meyer *et al.* (2003) improved the colour shading scheme which was proposed by Friendly (1994). In this extension, positive and negative residuals within the interval (–2, 2) are coded by white rectangles. The residuals exceeding −2 and 2 but still within the region (–4, 4) are shaded light red and light blue respectively. Furthermore, the residuals exceeding the limits of −4 and 4 are shaded in red and blue respectively. However, for this paper the colour coding system was changed in order to obtain monochrome figures and it is stated below.

▓ Pattern 1 for Red
░ Pattern 2 for Light Red
□ Pattern 3 for White
▤ Pattern 4 for Light Blue
■ Pattern 5 for Blue

Since Pearson residuals are approximately standard normal, the highlighted cells between (−2, −4) and (2, 4) contain residuals individually significant at approximately 5% level while the cells out of the (−4, 4) region are individually significant at approximately 0.01% level.

The applications of mosaic plots mentioned above can be carried out using the 'VCD' (Meyer *et al*. 2003) package in an R environment.

Univariate test

Since the data are of a hierarchical nature, the impact of the cluster level variable also should be taken into account when doing univariate analysis. Therefore, the most commonly used tests such as Pearson's Chi squared tests are not applicable to this scenario. Due to the stratified nature of the data and the presence of correlation between individuals within the clusters, Generalized Cochran Mantel–Haenszel test was proposed (Zhang & Boos 1997). Therefore, the univariate phase relied on the results of GCMH test which was used to identify the individual significance of both water and sanitation for the mortality rates of the two diseases. The R macro which has been developed by De Silva & Sooriyarachchi (2012) has been used for this. The theory behind this test is well explained in Ranatunga & Sooriyarachchi (2017).

Advanced methods

In this section a simple explanation is given of the advanced methodology and modeling. The more statistically advanced reader is referred to the paper by Ranatunga & Sooriyarachchi (2017) for a more technical discussion.

For the advanced analysis, multivariate multilevel binomial probit regression model was adopted. The software used was MLwiN version 2.10. Though the logit link is the most commonly used link function in practice, in MLwiN the multivariate model for binary responses has been developed only for the probitlink (Browne 2009).

Forward selection procedure and backward elimination procedure along with the Wald statistic and deviance information criteria were specifically used for the model building purpose (Agresti 2002). In order to assess the model adequacy, caterpillar plots and normal probability plots were used.

In the presence of some ordinal categorical responses, as cofactors in the model, MLwiN 2.10 defines the lowest category as the base level. It is more accurate, however, to change the base level according to the nature of the relationship between the response and the cofactor.

In this dataset, improved drinking water sources and improved sanitation may lead to a decrease in the incidence of all diseases. Therefore, it can be seen that the incidence of diseases is increasing when the level of the quality of water and sanitation are decreasing. Therefore, it would be more meaningful and practicable to get the highest level as the reference for both water and sanitation.

Table 1 represents the categorization of water and sanitation used in the analysis phase: 'water' and 'sanitation' represent 'The population using improved drinking water sources (%)' and 'The population using improved sanitation (%)' respectively. The three categories are low, moderate and high quality.

RESULTS AND DISCUSSION

Graphical phase

This part is a simple but crude analysis as the hierarchical nature of the data was not taken into account under this phase.

Table 1 | Categorization of water and sanitation

Cofactors	Category (%)	Coding	Quality
Water	<88	Water 1	Low
	88–98	Water 2	Moderate
	>98	Water 3	High
Sanitation	<40	Sanitation 1	Low
	40–80	Sanitation 2	Moderate
	>80	Sanitation 3	High

Water3 and Sanitation3 were used as base categories.

Here, the analyses were performed using mosaic plots in Figures 1 and 2 which acted solely as visualization tools. This was then followed by a Univariate analysis to progress forward.

Mosaic plots show that the Pearson residuals are highly significant at the 5% level for both diseases. It indicates a strong relationship between the two diseases and water. The groups of the lowest levels of the diseases having the

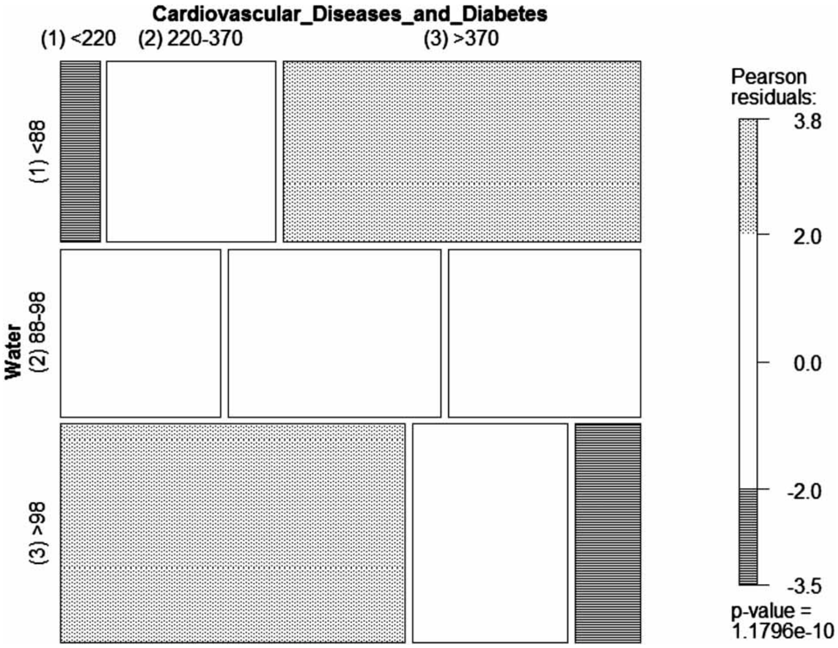

Figure 1 | Mosaic plot of CDD vs. water.

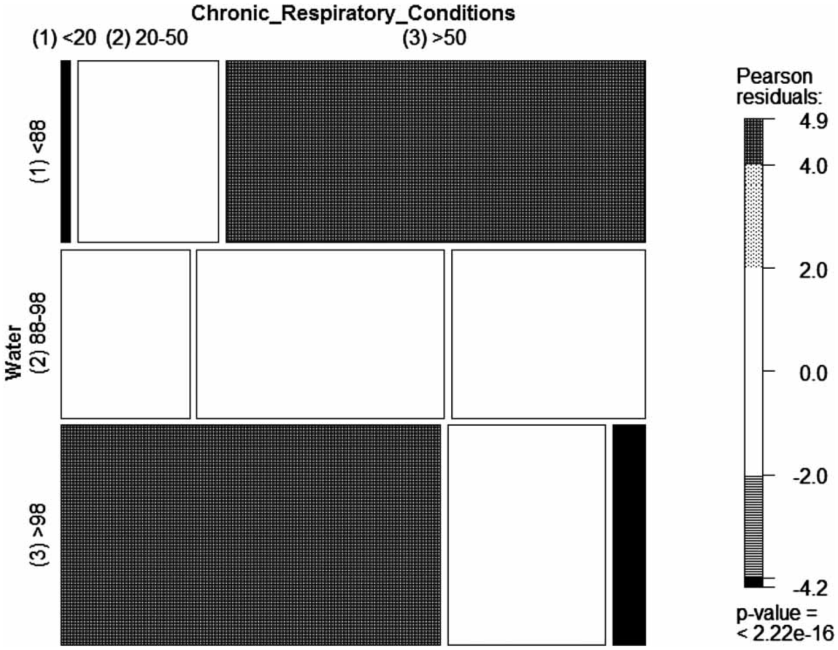

Figure 2 | Mosaic plots of CDD vs. water and CRC vs. water.

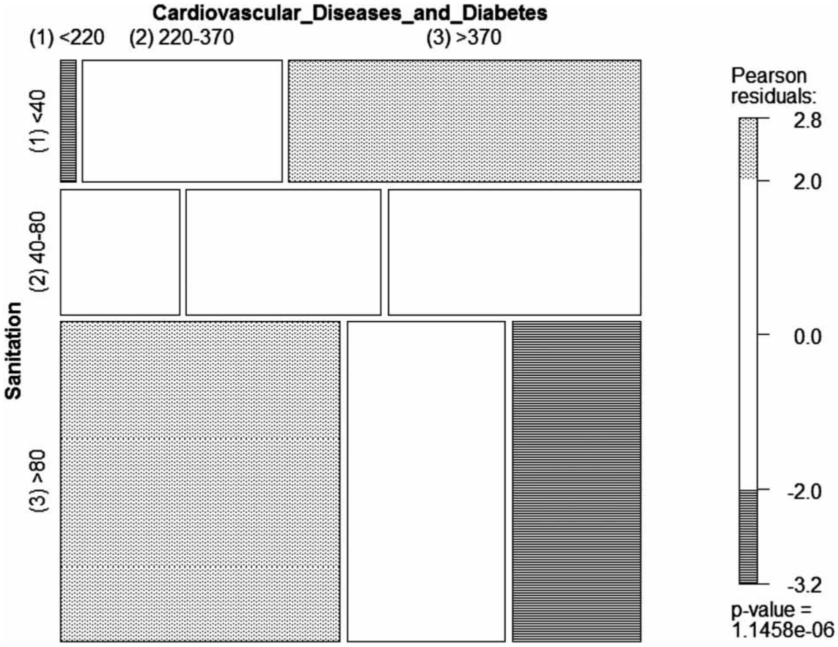

Figure 3 | Mosaic plots of CDD vs. sanitation.

highest level of water imply observed frequencies which are higher than that which would be expected under independence. The same conclusion can be drawn for the groups of the highest levels of diseases at the lowest level of water. Contrastingly, there are two groups with fewer numbers of observations which have lower observed frequencies than would be expected under independence. These two groups are shaded in Pattern 4 and 5 respectively for both diseases.

Figure 3 shows the mosaic plots of CDD versus sanitation. Sanitation also shows a same pattern of significance as water for the disease CDD. However, sanitation does not show any visually illustrated significance for the mortality rates of CRC. This may have happened due to ignoring the correlation structure which was actually present.

Univariate phase

Univariate analysis facilitates the examination of all the variables one-by-one, separately. Since the data is hierarchical in nature, the impact of the cluster level variable also should be taken into account. Therefore the most commonly used tests such as Pearson's Chi squared tests are not applicable to this scenario. Consequently, the GCMH test was implemented (Zhang & Boos 1997).

When modeling the multivariate (multiple) response for univariate analysis, CDD and CRC were used and a single response variable was constructed to represent the combined levels of the two response variables, as presented in Table 2. This examines the interaction between the two diseases. The four levels in Table 2 consist of countries having lower levels of deaths for both CDD and CRC, lower level of CDD deaths but upper level of CRC deaths, upper level of CDD deaths but lower level of CRC deaths and upper level of both CDD and CRC deaths.

As this is a univariate analysis, a liberal significance of 20% level was taken. (Collett 2002; De Silva & Sooriyarachchi 2012). This significance level can be increased because more stringent significance levels can lead to the exclusion

Table 2 | Description of combined levels of diseases

CDD	CRC	Coding for composite outcomes
1	1	1
1	2	2
2	1	3
2	2	4

of potentially useful predictor variables. Table 3 shows the results of the univariate analysis for the composite outcome.

According to Table 3, both water and sanitation are significant at the 20% level of significance for the composite outcome. Furthermore, other risk factors also show significant relationships to the combination of both diseases.

Advanced analysis

As CDD and CRC are highly correlated (0.680), it is more efficient to model these diseases in a multivariate framework using multilevel multiariate models. Also CDD and CRC have many explanatory variables in common. Tables 4 and 5 depict the results of the fitted model for CDD and CRC respectively. Though a uniariate analysis was carried out and variables selected at this stage, as only a few explanatory variables were available all of these were used in the modeling phase. The model that is fitted is a multilevel multivariate binary model with the probit link function. In this case the parameter estimates are interpreted using probability diferences. As the responses are binary and not normal, means (averages) are not used to interpret the model parameters but probability differences are used in that place (Agresti 2002). The interpretation of probability differences are as follows. For the first parameter, when all continuous variables are at their average values and when sanitation is at its base level the difference in the probability of mortality on CDD between low quality water and high quality water is 0.6478. This shows that high water quality reduces the mortality of CDD when compared to low water quality. Other parameters are interpreted similarly.

Table 3 | Test results for composite variable of two diseases vs. risk factors

Risk factors	Tp (GCMH test statistic)	Degrees of freedom (DF)	*P* value
Water	27.201	6	<0.001
Sanitation	19.810	6	0.003
Solid fuel	31.403	6	<0.001
Blood glucose	14.572	6	0.024
Blood pressure	21.195	6	0.002
Obese	19.385	6	0.004
Alcohol	15.797	9	0.071
Smoking	16.869	9	0.051

Table 4 | Differences for CDD

Term	Probability difference
Water 1 – Low quality (when continuous variables are at average and sanitation = base level)	0.6478
Water 2 – Moderate quality (when continuous variables are at average and sanitation = base level)	0.3106
Sanitation 1 – Low quality (when continuous variables are at average and water = base level)	Not significant
Sanitation 2 – Moderate quality (when continuous variables are at average and water = base level)	Not significant

Base level (Level 3)–high quality.
The meaning of low, moderate and high quality water and sanitation levels are given in Table 1.

Table 5 | Probability differences for CRC

Term	Probability difference
Water 1 – Low quality (when continuous variables are at average and sanitation = base level)	0.5745
Water 2 – Moderate quality (when continuous variables are at average and sanitation = base level)	0.2344
Sanitation 1 – Low quality (when continuous variables are at average and water = base level)	0.4911
Sanitation 2 – Moderate quality (when continuous variables are at average and water = base level)	0.2067

Difference cannot be calculated between water level 1 and water level 2 and sanitation level 1 and sanitation level 2 as these are not continuous variables but ordinal categorical variables.

Effect of water

In the final model, water acts as a common coefficient and it has three levels. Water 1 (low quality) takes a probability value of 0.6478 and it implies that the probability of being the higher mortality group of CDD is 0.6478 higher when water is at level 1 (low quality) when compared to when water is at level 3 (high quality, >98% of population using improved drinking water sources) while all the other continuous variables are taken at average and sanitation is taken at the base level. Similarly, when water is at level 2 (moderate quality), it takes a probability value of 0.3106. Furthermore, it can be seen that the probability of being in the higher mortality group of CDD for water 1 (low quality) is approximately twice as high as water 2 (moderate quality). Therefore, when the usage of improved drinking water sources decreases, the probability of being in the higher group of CDD increases.

Past evidence is also available to prove this relationship. A study conducted in Newfoundland has shown that the proportion of mortality rates for CDD was higher in the soft water areas than hard water areas (Fodor *et al.* 1973). It further showed that there was a macro geography variation for CDD. Those findings tally with the findings in this research since CDD also shows a continent level variation (Ranathunga & Sooriyarachchi 2017).

Effect of sanitation

For sanitation, it has separate coefficients. However, both levels of sanitation do not have a significant impact for the probability of being in the higher mortality group of CDD when compared to the sanitation level 3 (high quality, >80% of population using improved sanitation).

Effect of water

Similar to CDD, the probability of being in the higher mortality group of CRC is 0.5745 more when water is at level 1 (low quality) when compared to when water is at level 3 (high quality, >98% of population using improved drinking water sources) while all the other continuous variables are taken at the average and the sanitation is taken at the base level. In the same way, the probability of being in the higher mortality group of CRC is 0.2344 higher when water is at level 2 (moderate quality). Furthermore, it can be seen that the probability of being in the higher mortality group of CRC for water 1 (low quality) is approximately twice as high as water 2 (moderate quality). Therefore, when the usage of improved drinking water sources decreases, the probability of being in the higher group of CRC increases.

According to past evidence there is also a suggestion that water has an impact on the CRC. The US Environmental Protection Agency has shown that heavy rainfall events cause storm water overflow that may contaminate water bodies used for drinking with other bacteria. It may cause illnesses including ear, nose, and throat infections (Climate Impacts on Human Health n.d.).

Moreover, it is interesting to note that when the usage of improved water sources decrease from level 2 to 1, the probability of getting higher risk is doubled for both CDD and CRC. However, the impact of water for CDD is higher than that for CRC.

Effect of sanitation

The probability of being in the higher mortality group of CRC is 0.4911 more when sanitation is at level 1 (low quality) and 0.2067 more when sanitation is at level 2 (moderate quality) when compared to when sanitation is at level 3 (high quality, >80% of population using improved sanitation) while all the other continuous variables are taken at average and the water is taken at the base level. Furthermore, it can be seen that the probability of being in the higher group of CRC for sanitation 1 (low quality) is approximately twice as high as sanitation 2 (moderate quality). Therefore, it can be seen that when the usage of improved sanitation sources decreases, the probability of being in the higher group of CRC increases.

When considering the risk factors for the CDD and CRC, most articles stated only the major well known medical risk factors such as obesity, smoking, alcohol, blood pressure and blood glucose, etc., for the vulnerability of diseases (World Heart Federation n.d.). Therefore, when identifying the relationship between risk factors and diseases, the majority of the studies focus on the medical impact instead of the environmental impact.

According to an article from the Bill & Melinda Gates Foundation, nearly 40% of the world's population are suffering from inadequate sanitation facilities and most of them represent developing countries (Water, Sanitation & Hygiene, Strategy Overview n.d.). The penalties of these can be directly impacted to human health in various ways. Not only the developing countries, but also the rich countries such as USA, Japan, etc., must consider supplying safe water and sanitation for all. According to the United Nations expert, there is still no access to safe water and sanitation for homeless due to discrimination (UN News Centre 2011). Solving this challenge would require carrying out more studies worldwide and that would contribute to social, environmental and economic development.

Findings of this study have also proven that the usage of poor quality water and sanitation sources also have an impact on the diseases CDD and CRC after adjusting for these well-known direct risk factors. According to Briggs (2003), unsafe water, poor sanitation and poor hygiene are seen to be one of the major sources of exposure for these types of diseases.

Table 6 shows the estimates (95% credible intervals) of the continent level variances from the fitted final model.

These intervals are Bayesian credible intervals and therefore *P*-values are not available, unlike in the frequentist approach. The fact that the Bayesian credible interval for the continent level variance does not include zero indicates that the continent level variance is significant for the joint distribution of CDD and CRC.

In order to check the suitability of applying the multilevel concept to the multivariate model, it is important to check the significance of the continent level variance. If the continent level variance is zero then it is not meaningful to conduct a multilevel analysis since it is equivalent to the single level case.

H_0: Unexplained continent level variance is zero

versus

H_1: Unexplained continent level variance is not zero

As zero is not included in the 95% credible interval, H_0 is rejected and it is concluded that the continent level variance is significant, implying that the multilevel approach for the multivariate context is suitable.

Table 6 | Estimate (with 95% credible interval) of the continent level variance for multivariate multilevel probit model

Parameter	Estimate (95% credible interval)
Unexplained continent level variance	1.425(0.254, 5.109)

After fitting the model we then then focused on examining the model adequacy techniques to check whether the fitted models are adequate or not. In order to do that, caterpillar plots and normal probability plots in Figure 4 were used.

According to the caterpillar plot (first plot), four residuals do not contain zeros in their 95% confidence bands. They imply significant differences from the overall mean predicted by the fixed effects from the model. These four continents are North America, Europe, Asia and Oceania. Moreover, it can be seen that Asia and Oceania show

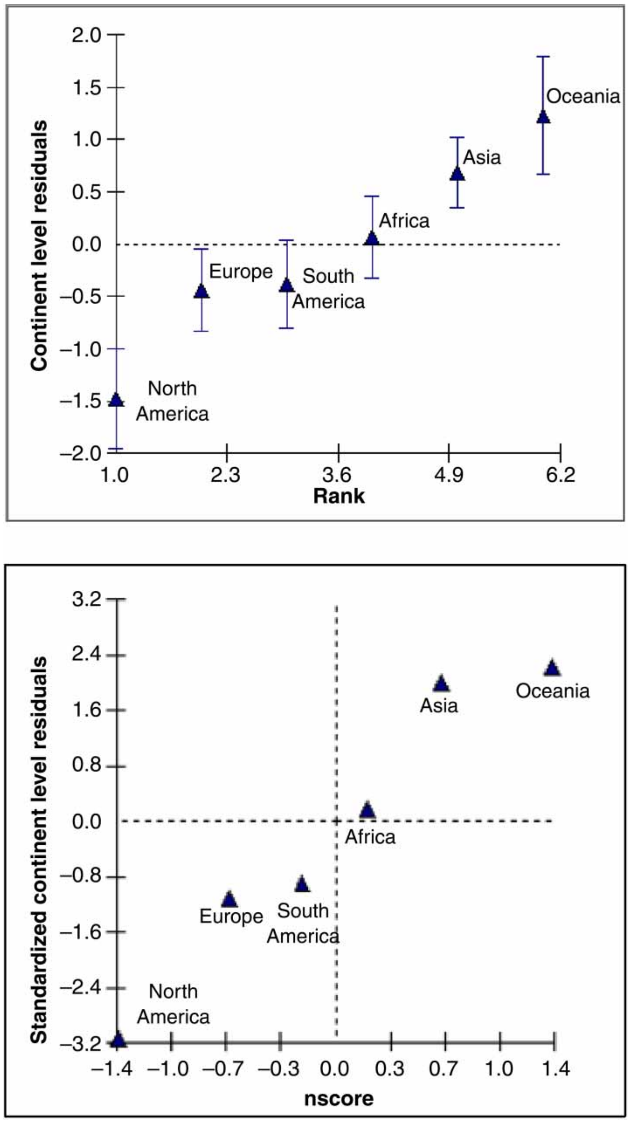

Figure 4 | Caterpillar plot and normal plot for multivariate multilevel probit model.

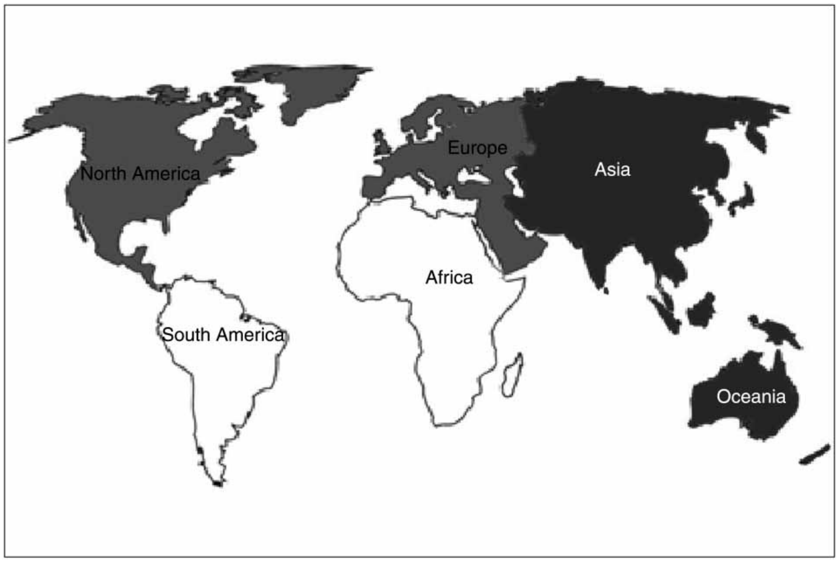

Figure 5 | Continent variation for CDD and CRC.

positive residual deviations while North America and Europe show negative deviations. Therefore, it is possible to conclude that these four continents contribute to a high continent effect for both CDD and CRC when taken together. The normal plot (second plot) suggests that the points are approximately through the 45° axis. Therefore, it can be concluded that the residuals are approximately normally distributed (Anderson Darling Test, Ranathunga & Sooriyarachchi 2017).

Figure 5 illustrates the continent variation for CDD and CRC more clearly. and is explained in Ranatunga & Sooriyarachchi (2017).

The continents which have a low risk of CDD and CRC are symbolized by the ash colour and the continents which have a high risk of CDD and CRC are symbolized in black.

CONCLUSIONS

When analyzing risk factors for diseases, most studies usually consider only the well-known medical risk factors rather than environmental risk factors. A paper to the former effect is in line for publication (Ranatunga & Sooriyarachchi 2017). This paper looks at the later problem. Though less consideration is given to factors such as water and sanitation, for the diseases CDD and CRC, these factors can also be shown as contributory risk factors responsible for the worldwide mortality rates of these diseases after adjusting for the more well-known factors. There is little literature on this subect and the reader is referred to Belue *et al.* (2009) and Sengupta (2013) for CDD and www.who.int/mediacentre/news/releases/2016/deaths-attributable-to-unhealthy-environments/en/, Briggs (2003) for CRR.

REFERENCES

Agresti, A. 2002 *Categorical Data Analysis*, 2nd edn. John Wiley & Sons, USA.

BeLue, R., Okoror, T. A., Iwelunmor, J., Taylor, K. D., Degboe, A. N., Agyemang, C. & Ogedegbe, G. 2009 An overview of cardiovascular risk factor burden in sub-Saharan African countries: a socio-cultural perspective. *Global. Health* **5**, 10.

Briggs, D. 2003 Environmental pollution and the global burden of disease. *Br. Med. Bull.* **68** (1), 1–24.

Browne, W. J. 2009 *MCMC Estimation in MLwiN v2.10.* Centre for Multilevel Modeling, University of Bristol, Bristol, UK.

Climate Impacts on Human Health n.d. World Heart Federation. Available from: www.epa.gov/climatechange/impacts-adaptation/health.html (accessed 12 April 2014).

Collett, D. 2002 *Modelling Binary Data,* 2nd edn. CRC Texts in Statistical Science, Chapman and Hall, UK.

De Silva, D. & Sooriyarachchi, M. 2012 Generalized Cochran Mantel Haenszel test for multilevel correlated categorical data: an algorithm and R function. *J. Nat. Sci. Found. Sri Lanka* **40** (2), 137–148.

Fodor, J. G., Pfeiffer, C. J. & Papezik, V. S. 1973 Relationship of drinking water quality (hardness-softness) to cardiovascular mortality in Newfoundland. *Can. Med. Assoc. J.* **108** (11), 1369–1373.

Friendly, M. 1994 Mosaic displays for multi-way contingency tables. *J. Am. Stat. Assoc.* **89**, 190–200.

Fuster, V. & Kelly, B. B. 2010 *Promoting Cardiovascular Health in the Developing World: A Critical Challenge to Achieve Global Health*. National Academies Press, Washington, DC.

Go, A. S., Mozaffarian, D., Roger, V. L., Benjamin, E. J., Berry, J. D., Blaha, M. J., Dai, S., Ford, E. S., Fox, C. S., Franco, S., Fullerton, H. J., Gillespie, C., Hailpern, S. M., Heit, J. A., Howard, V. J., Huffman, M. D., Judd, S. E., Kissela, B. M., Kittner, S. J., Lackland, D. T., Lichtman, J. H., Lisabeth, L. D., Mackey, R. H., Magid, D. J., Marcus, G. M., Marelli, A., Matchar, D. B., McGuire, D. K., Mohler, E. R. 3rd, Moy, C. S., Mussolino, M. E., Neumar, R. W., Nichol, G., Pandey, D. K., Paynter, N. P., Reeves, M. J., Sorlie, P. D., Stein, J., Towfighi, A, Turan, T. N, Virani, S. S, Wong, N. D, Woo, D, Turner, M. B. & American Heart Association Statistics Committee and Stroke Statistics Subcommittee 2014 Heart disease and stroke statistics. *Circulation* **129** (3), e28–e292.

Meyer, D., Zeileis, A. & Hornik, K. 2003 Visualizing independence using extended association plots. *Proceedings of the 3rd International Workshop on Distributed Statistical Computing* (K. Hornik, F. Leisch & A. Zeileis, eds). http://www.R-project.org/conferences/ DSC-2003/Proceedings/.

Ranathunga, K. N. O. & Sooriyarachchi, M. R. 2017 Multivariate multilevel modeling of age related diseases. *J. Mod. Appl. Stat. Meth.* **16** (1), 498–517.

Sengupta, P. 2013 Potential health impacts of hard water. *Int. J. Prevent. Med.* **4** (8), 866–875.

UN News Centre 2011 Rich countries fall short on providing safe water, sanitation for all. Available from: www.un.org/apps/news/story.asp?NewsID=39590#.WBBMxCT6OZR (accessed 27 October 2016). UN Expert, USA.

Water, Sanitation & Hygiene, Strategy Overview n.d. Available from: www.gatesfoundation.org/What-We-Do/Global-Development/Water-Sanitation-and-Hygiene (accessed 27 October 2016).

World Health Statistics 2013 World Health Organization, Geneva. Available from: http://apps.who.int/iris/bitstream/10665/81965/1/9789241564588_eng.pdf (retrieved 22 June 2017).

World Heart Federation Cardiovascular Disease Risk Factors n.d. Available from: www.world-heart-federation.org/cardiovascular-health/cardiovascular-disease-risk-factors/physical-inactivity/ (accessed 27 October 2016).

Zhang, J. & Boos, D. D. 1997 Mantel-Haenszel test statistics for correlated binary data. *Biometrics* **53**, 1185–1198.

First received 30 July 2016; accepted in revised form 8 June 2017. Available online 14 July 2017