

METHODOLOGY

Open Access



Within- and between-group regression for improving the robustness of causal claims in cross-sectional analysis

Bernd Genser^{1,2}, Carlos A. Teles^{2,3}, Mauricio L. Barreto^{2,3} and Joachim E. Fischer^{1*}

Abstract

Background: A major objective of environmental epidemiology is to elucidate exposure-health outcome associations. To increase the variance of observed exposure concentrations, researchers recruit individuals from different geographic areas. The common analytical approach uses multilevel analysis to estimate individual-level associations adjusted for individual and area covariates. However, in cross-sectional data this approach does not differentiate between residual confounding at the individual level and at the area level. An approach allowing researchers to distinguish between within-group effects and between-group effects would improve the robustness of causal claims.

Methods: We applied an extended multilevel approach to a large cross-sectional study aimed to elucidate the hypothesized link between drinking water pollution from perfluorooctanoic acid (PFOA) and plasma levels of C-reactive protein (CRP) or lymphocyte counts. Using within- and between-group regression of the individual PFOA serum concentrations, we partitioned the total effect into a within- and between-group effect by including the aggregated group average of the individual exposure concentrations as an additional predictor variable.

Results: For both biomarkers, we observed a strong overall association with PFOA blood levels. However, for lymphocyte counts the extended multilevel approach revealed the absence of a between-group effect, suggesting that most of the observed total effect was due to individual level confounding. In contrast, for CRP we found consistent between- and within-group effects, which corroborates the causal claim for the association between PFOA blood levels and CRP.

Conclusion: Between- and within-group regression modelling augments cross-sectional analysis of epidemiological data by supporting the unmasking of non-causal associations arising from hidden confounding at different levels. In the application example presented in this paper, the approach suggested individual confounding as a probable explanation for the first observed association and strengthened the robustness of the causal claim for the second one.

Keywords: Causal claims, Cross-sectional studies, Multilevel modelling, Ecological fallacy, Ecological inference

Background

An important issue in environmental epidemiology is the robustness of causal claims linking exposures to adverse health outcomes. Strong support for causality arises if dose–response relationships between exposures and outcome at the individual level can be demonstrated. However, exposures (e.g. air pollution, persistent chemicals) are often spatially correlated and vary relatively little

within single geographic regions. Thus, to obtain sufficient variance of exposure concentrations researchers often recruit individuals from different geographic areas. Compared to measuring exposure directly at the individual level, such assessment of exposure at the area level offers the advantage of lower costs. This approach, referred to as ecological studies or ecological inference, has emerged as an avenue for studying exposure–health associations at the macro-level [1–4]. However, if causal claims are a research objective [5–8], investigators must focus on individual-level exposure–outcome analysis (“biologic inference”) [1]. In most cases individual-level associations cannot be deduced from group-level associations, a phenomenon

* Correspondence: Joachim.Fischer@medma.uni-heidelberg.de

¹Mannheim Institute of Public Health, Social and Preventive Medicine, University of Heidelberg, Ludolf-Krehl-Strasse 7-11, Mannheim 68167, Germany

Full list of author information is available at the end of the article

referred to as ecological fallacy or cross-level bias. Cross-level bias arises from “context effects” (i.e. effects of neighbourhoods on the individual-level exposure-outcome relationships) or from confounding or biases arising differentially at the individual or group level [9–11].

Simple statistical analysis of individual-level data also fails to offer straightforward solutions due to the hierarchical clustering of determinants on different levels, e.g. ecological factors such as environmental exposure concentrations and individual factors such as daily intake, absorption and excretion. To address these issues, researchers increasingly employ hierarchical modelling techniques (multilevel modelling), which allow covariates at different levels [12–14] to be included. However, in cross-sectional data analysis, even these advanced approaches have limitations with regard to the robustness of causal claims. The resulting estimate from such multilevel analysis is a single coefficient for the exposure-outcome relationship that fails to disentangle within-group effects from between-group effects.

The present work describes an approach, which partitions within- and between-group relations. Originally developed in the social sciences, the methodology is *within- and between-group regression* (WBGR) [15]. For example, the average intelligence quotient (IQ) of a school class commonly affects the individual-level association between the IQ and learning performance of students. However, this WBGR approach is rarely applied in environmental epidemiology [16].

The aim of our paper is to generalise the WBGR approach for epidemiological studies where individuals are recruited from different geographical areas and where environmental exposure varies between areas. In these studies, the variation of exposure within and between areas is affected by different factors. Between-area variation results from the effect of area-specific variables (i.e. the magnitude of environmental exposure concentration). Thus, exposure-outcome relations analysed within areas (individual level) and between areas (group level) may yield different results because the relations are prone to different sources of bias or to the presence of a context effect. The latter implies that the average level of a given group affects the individuals’ within-group relation (e.g. the effect of neighbourhoods on individual-level exposure-outcome relations). In the present paper, we illustrate this WBGR approach using a dataset from a cross-sectional study, which primarily investigates whether serum concentrations of the chemical substance perfluorooctanoic acid (PFOA) affected different health-related outcomes.

Methods

We briefly introduce the basic concepts of WBGR for the applied researcher; mathematical details of the procedure are described in the appendix.

The conventional multilevel approach for clustered data employs equation (1)

$$Y_{ij} = \alpha_{00} + \beta_{10} x_{ij} + U_{0j} + R_{ij} \quad (1)$$

which incorporates an intercept α_{00} , a fixed effect β_{10} denoting the total effect observed at the individual level and a random effect of the group U_{0j} ; R_{ij} denotes the residual error term at the individual level within group j . Researchers usually include additional covariates at both individual and group levels to minimize the residual error. If the observed total effect β_{10} (TE) has exactly the same magnitude as the association observed within areas, the within-group effect (WGE), and as the association observed between area-average values, the between-group effects (BGE), the analysis could safely stop here, as there is little evidence for confounding on the individual or area level. However, equation (1) does not provide this information.

Within- and between-group regression

The first step of WBGR is to elucidate whether there is homogeneity of the different effect estimators described above, i.e. whether $TE = WGE = BGE$. We achieve this by introducing an additional term into equation (1), namely, the group average of the individual-level exposure concentrations (equation 2):

$$Y_{ij} = \alpha_{00} + \beta_{10} x_{ij} + \beta_{01} \bar{x}_j + U_{0j} + R_{ij} \quad (2)$$

In equation (2) the parameter β_{01} quantifies the difference between WGE and BGE; the derivation of this relationship is briefly described in the appendix and more mathematical details are described elsewhere [17]. Rejecting the null hypothesis of $\beta_{01} = 0$ implies an effect of the exposure concentration averaged at the group level beyond the WGE. This test procedure is known in econometrics as the *Hausman test* [18]. For $\beta_{01} > 0$, the BGE is larger than the WGE, while for $\beta_{01} < 0$ the BGE is smaller than the WGE (see section below for interpretation).

If $\beta_{01} \neq 0$, BGE and WGE should be explicitly estimated by fitting a second model with a slightly different parameterisation. This we achieve by including the deviation of the individual exposure from the group average ($x_{ij} - \bar{x}_j$) as a predictor, resulting in equation (3):

$$Y_{ij} = \tilde{\alpha}_{00} + \tilde{\beta}_{10} (x_{ij} - \bar{x}_j) + \tilde{\beta}_{01} \bar{x}_j + U_{0j} + R_{ij} \quad (3)$$

now explicitly estimating *BGE* ($\tilde{\beta}_{01} = \beta_{10} + \beta_{01}$) and *WGE* ($\tilde{\beta}_{10} = \beta_{10}$).

The extension of WBGR to more complex study designs with more than two levels of clustering is straightforward, as illustrated in the following example. A multi-centre study recruited students from different schools. To disentangle within-class effects, between-class effects and between-school effects, we introduce

two additional random effects at the class and at the school level with their respective aggregated exposure levels. Expanding equation (3) allows the size of the between-class and between-school effect to be compared to exposure-outcome associations at the individual level. Mathematical details and applications of three- or higher order multilevel modelling are described elsewhere [17].

Adjustment for residual within-group clustering

In multilevel modelling, random effects are the preferred approach to address residual within-group clustering [13, 19]. Conceptually, the random variable represents the effects of all unobserved determinants at the group level. Presence of clustering should be tested by a hypothesis test based on an estimate of the variance of the random effect. Since random-effect modelling has stringent data assumptions (e.g. sufficient number of groups, distributional assumptions of the random effect), robust alternatives such as generalised estimating equations (GEE) or robust variance estimation are often preferable [20–22].

Interpretation of WGE and BGE

The simplification above does not account for clustering of individuals within groups, which is examined by determining the intra-class correlation coefficient (ICC). In the classical example of examining the relationship between IQ and academic achievement in students clustered within classes, WGE is the impact of the students' IQs on the performance score averaged across classes. By contrast, the BGE expresses the effect of the group mean IQ on the group mean performance scores. The overall exposure-outcome relation (TE) estimated by equation (1) is a weighted average of the underlying within- and between-group relations with weights proportional to the (ICC) [17]:

$$TE = BGE * ICC + WGE * (1-ICC). \quad (4)$$

Equation (4) implies that the TE must always be between BGE and WGE. If clustering is substantial (e.g. ICC > 0.7), the TE will be very close to the BGE. However, if there is only little clustering (e.g. ICC < 0.3), the TE will be close to the WGE. The larger the clustering, the better the BGE represents the TE.

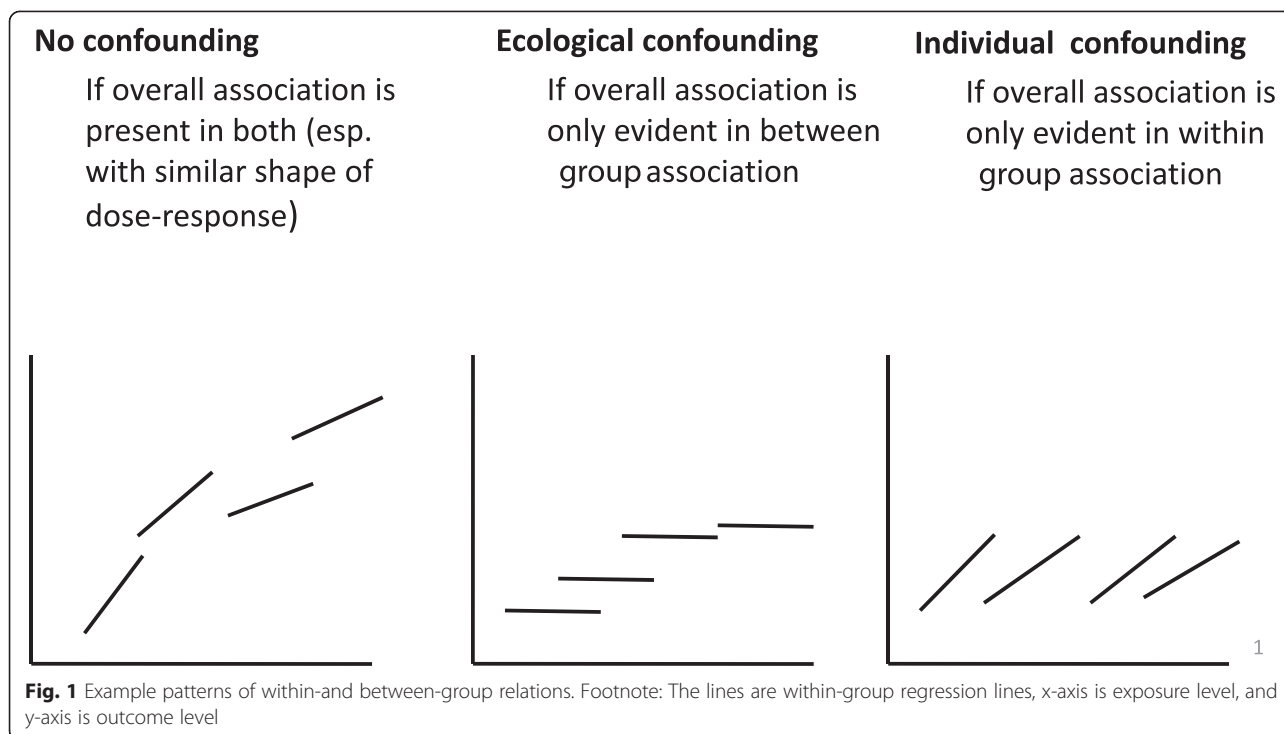
WBGR and context effects in environmental epidemiology

True context effects are very common in research on psychosocial determinants of health and have been acknowledged in the social sciences since Émile Durkheim's seminal work on regional differences in suicide rates more than a century ago [23]. In contrast, in most environmental epidemiological applications, true context effects are rare. A true context effect in the present

example of the C8 health project would imply that the biological effect of PFOA on an individual's health is affected by the average district concentration of PFOA serum exposure beyond the individual PFOA serum concentration. This is rather unlikely, since an individual's PFOA serum concentration results from the individual's consumption of polluted water; a person drinking solely bottled water would not have any PFOA exposure at all. Because the effect of the exposure at the macro-level (PFOA concentration in the water district) is reflected in the PFOA exposure in the serum, we can exclude here another causal pathway from the average district level to the health outcomes. Thus, if true context effects are unlikely in terms of biological reasoning but a WBGR analysis reveals the presence of considerable between-group associations (as in our example), at least this part of the total effect cannot be explained by biases or confounding at the individual level. To the extent that we can exclude confounding at the group level, this finding may be interpreted to improve causal claims as illustrated below. In contrast, if the between-group effect estimate is negligible in comparison to the within-group effect or the individual level outcome-exposure association, the presence of bias and non-causality is more likely (Fig. 1) [9].

An application example – the C8 Health Study

To illustrate the WBGR approach, we analyse data from a large cross-sectional study, the *C8 Health Project*. The study was approved by the *London School of Hygiene and Tropical Medicine Ethics Committee* and is one of the *C8 Science Panel studies*; details of the study design are described elsewhere [24]. Briefly, the aim of the study was to elucidate the possible association between the toxic pollutant PFOA and intermediate health outcomes (biomarkers) and clinical outcomes in 69,030 people living in different water districts exposed to environmental pollution by a chemical plant emitting PFOA used in the manufacturing of fluoropolymers. Eligible study participants were recruited between August 2005 and August 2006 in the states of Ohio and West Virginia, USA. Individuals were eligible to participate if they had consumed water for at least one year between 1950 and 2004 while living, working or going to school in one of the six water districts, in an area of private water sources or in areas of documented PFOA pollution (participation rate 80 %). A separate analysis identified residence in one of the contaminated water districts as the strongest predictor of individual PFOA serum concentration [25]. As supporting the causal claim of linking PFOA pollution to health outcomes was a major objective, researchers decided to use the individual's PFOA serum concentrations as the exposure measure and a variety of health measures as outcomes, including an array of intermediary biomarkers [26–30]. For the



present application example we selected a subpopulation of the *C8 Health Project* study population consisting of 25,817 adults (> 18 years) that were stable residents of six different water districts with different PFOA exposure concentrations in the drinking water.

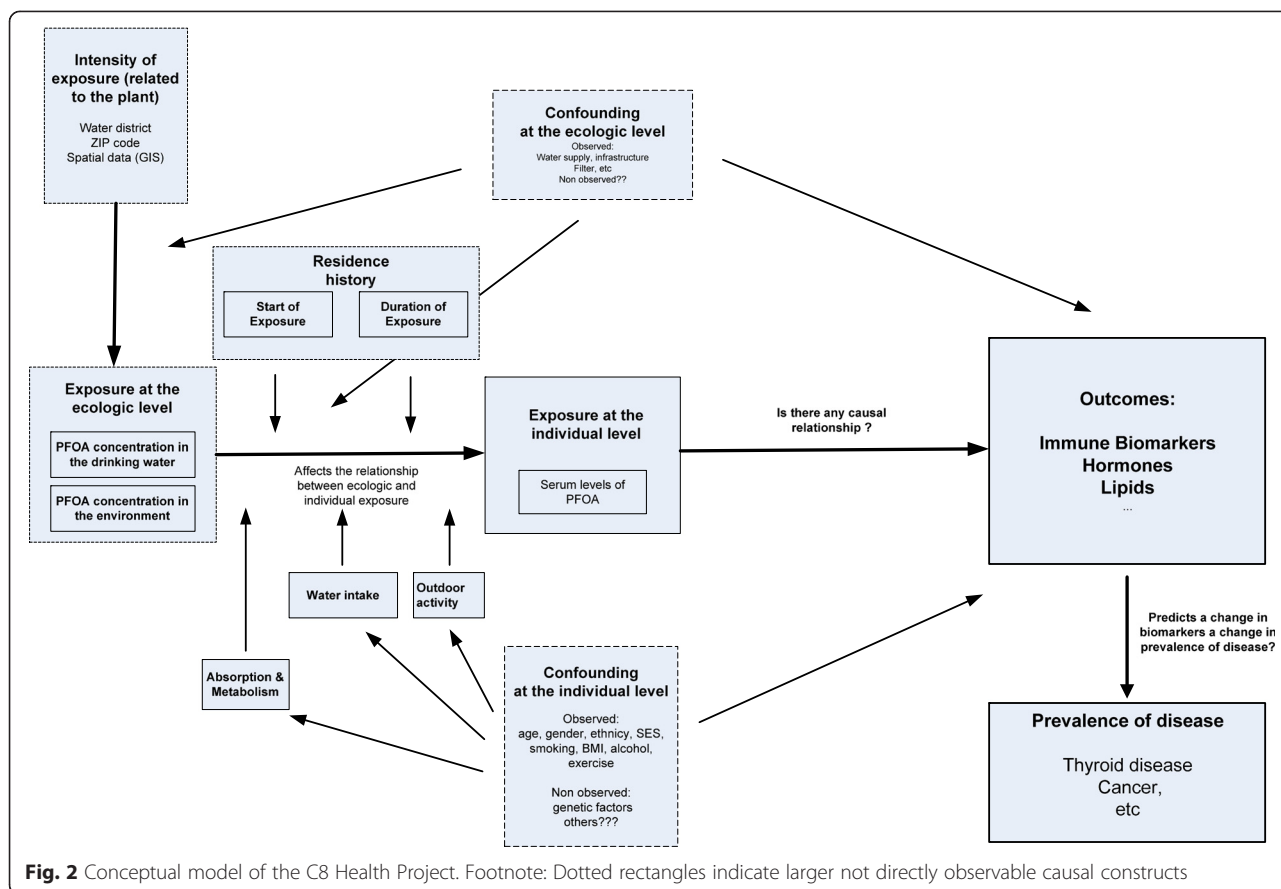
Individual PFOA serum concentrations ensued primarily from water intake from a contaminated drinking water supply. However, an individual's serum concentrations can ensue both from regional characteristics (i.e. between-group differences in drinking water concentration of PFOA) and from multiple individual factors affecting the bioaccumulation of PFOA (i.e. within-group differences, such as genetic factors, amount of daily water intake, duration of residence in the area). Thus, the exposure was the result of a complex causal chain of exposure determinants acting on different levels (Fig. 2). As the perfect district level exposure measurement (i.e. drinking water concentrations of PFOA at time of water consumption) was not available, the individual PFOA serum concentration reflecting the integrated bioaccumulation over time had to be used as best proxy.

Other research has used such proxies in the concept of instrumental variables [31–33]. An instrumental variable is an operationalisation of an intermediary variable within the causal chain leading from exposure to health outcomes. The key feature of the instrumental variable is its independence of any individual-confounding variable. For example, the association between LDL cholesterol (LDL-C) and cardiovascular disease may be confounded by

a myriad of individual behaviour variables, which lead both to elevated LDL-C levels and increased risk of cardiovascular disease. However, genetic variations leading to higher LDL-C levels are very unlikely to be confounded by behavioural variables acquired after birth. Likewise, in the present example, district-average PFOA serum concentrations may be employed as an instrumental variable for PFOA drinking water exposure, as long as individual confounders (e.g. genetic variation in metabolism or individual factors affecting water intake) are randomly distributed across water districts.

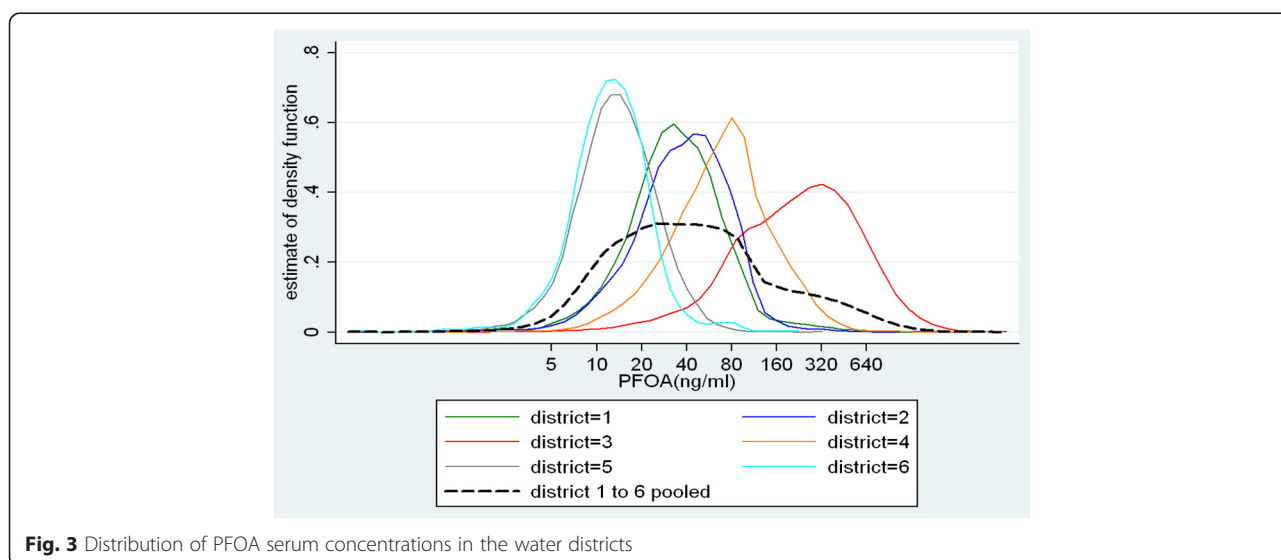
Applying WBGR to the data of the *C8 Health Project*

We conducted data analysis in several steps. First, we tested different linear models (total and stratified by district), assuming log-linear and log-log relations. All models were adjusted for potential individual-level confounding variables (age, gender, body mass index, frequency of exercise, alcohol consumption, month of measurement). Secondly, we visualised the relationship pattern by bar plots showing the fitted marginal means of outcomes vs. deciles of PFOA (total, stratified by district and aggregated by district). Third, we tested for heterogeneity of within-district slopes across districts by including interaction terms. Fourth, we assessed whether within- and between-district associations were different by using the WBGR approach, which incorporates the average PFOA level of each district as an additional explanatory variable. If we found heterogeneous BGE and



WGE, we fitted a second model using the deviance of individual PFOA exposure from the average PFOA exposure in the district. Finally, we tested for the presence of residual within-district clustering by estimating the variance of the random effect.

For the present illustration of the WBGR concept, we deliberately selected two biomarkers, which showed different patterns of within- and between-district relations: *lymphocyte count*, where we found evidence that the observed association is non-causal, and *C-reactive protein*



(CRP), where we found consistent within- and between-district associations.

Results

Serum concentrations of PFOA (measured in ng/ml) varied substantially across the six water districts (mean of district means = 89.8, SD = 118.0, min = 15.0, max = 324.2) and among the 25,817 individuals (mean = 102.9, SD = 187.7, min = 0.25, max = 4751.5). As expected, individual PFOA exposure concentrations clustered substantially within water districts (ICC = 46 %, Fig. 3). Likewise, we observed for CRP serum concentrations (measured in mg/l) substantial variance between districts (mean of district medians = 1.7, SD = 0.2, min = 1.5, max = 2.0) and between individuals (median = 2.4, iqr = 2.8, min = 0.2, max = 10.0). By contrast, lymphocyte counts showed very small variation between districts (mean of district means = 2.08, SD = 0.03, min = 2.04, max = 2.12) but strong variation between individuals (mean = 2.07, SD = 0.66, min = 0.20, max = 10.70). Calculation of ICC confirmed this observation: we found significant clustering of CRP concentrations within districts (ICC = 2 %) but absence of within-district clustering of lymphocyte count (ICC = 0 %).

Crude multilevel estimates of the association between PFOA serum concentrations, after adjusting for multiple confounders at the district and individual level, revealed significant overall associations ($p < 0.001$) with the two outcome variables (see total effects shown in Table 1). At this point, it was unclear whether these results were due to residual confounding not detected by a regular multilevel analytical approach.

Applying the strategy outlined in the methods section by introducing the district-average PFOA serum concentration as the term $\beta_{01} \bar{x}_j$ in equation (2) revealed β_{01} as different from 0 for both outcomes ($p < 0.01$). This variance suggested the need for further exploration using the WBGR approach. Estimating the BGE and the WGE using equation (3) revealed a significant BGE for CRP,

but no BGE for lymphocyte counts. Results are tabulated in Table 1 and illustrated in Figs. 4 and 5.

In summary, the R^2 contribution of PFOA was very small and quite similar for log-linear and log scales. For CRP, we observed consistent slopes within all districts (Fig. 5, panel b); slopes were less consistent for lymphocytes within district, showing a saturation effect in some districts (Fig. 4, panel b). If PFOA causes the observed relationship, we would also expect to see an association at the aggregated level (i.e. between districts shown in panel c). This was the case for CRP, where we also observed a clear trend on the aggregated level (Fig. 4c) but not for lymphocyte counts (Fig. 5c). Additionally, we found that the significant clustering of CRP concentrations within districts (ICC = 2 %) disappeared after adjusting for PFOA and other covariates, corroborating the hypothesis that part of the CRP variation is explained by heterogeneous PFOA concentrations between districts. Results from WBGR further corroborated this finding; WGE and BGE were of similar magnitude and statistically significant (Table 1). For lymphocytes, we observed heterogeneous WDE and BDE and heterogeneous WDE within districts, both indicating confounding and/or reverse causality.

Discussion

We presented WBGR as an approach for statistical analysis of clustered epidemiological data aimed at improving the robustness of causal claims in cross-sectional analysis. We illustrated the application of the approach to data from a large cross-sectional study with strong clustering of individual exposure concentrations (serum concentrations of PFOA) within water districts, which had been contaminated by the emissions from a chemical plant. By disentangling the exposure-outcome relations observed within- and between-groups, such as individuals living in a particular geographical area, the approach may reveal bias in estimates and indicate

Table 1 Results of within- and between-regression modelling of PFOA on immune biomarkers (the C8-Health project, N = 25 817)

Outcome	model	Between-district effect		Within-district effect		Total effect		Model fit				
		beta ^a	SE ^b	beta ^a	SE ^b	beta ^a	SE ^b	R ² _o ^c	R ² _d ^d	dR ² _e	R ² _w ^f	R ² _b ^g
CRP	log-linear	-0.232	0.0525	-0.113	0.0403	-0.157	0.0328	21.0 %	21.1 %	0.1 %	20.7 %	90.4 %
	log-log	-31.021	6.3289	-36.605	7.9959	-33.170	6.3668	21.0 %	21.2 %	0.2 %	20.8 %	92.8 %
Lymphocytes	log-linear	-0.012	0.0166	0.065	0.0127	0.036	0.0097	11.4 %	11.4 %	0.1 %	11.2 %	88.2 %
	log-log	-1.365	1.9997	23.091	2.5265	8.045	3.8012	11.4 %	11.5 %	0.1 %	11.4 %	88.4 %

^a: regression coefficient of PFOA * 1000

^b: standard error of regression coefficient * 1000

^c: coefficient of determination of a model only including covariates without PFOA

^d: coefficient of determination of a model additionally including PFOA

^e: difference in R² due to inclusion of PFOA

^f: coefficient of determination within districts

^g: coefficient of determination between districts

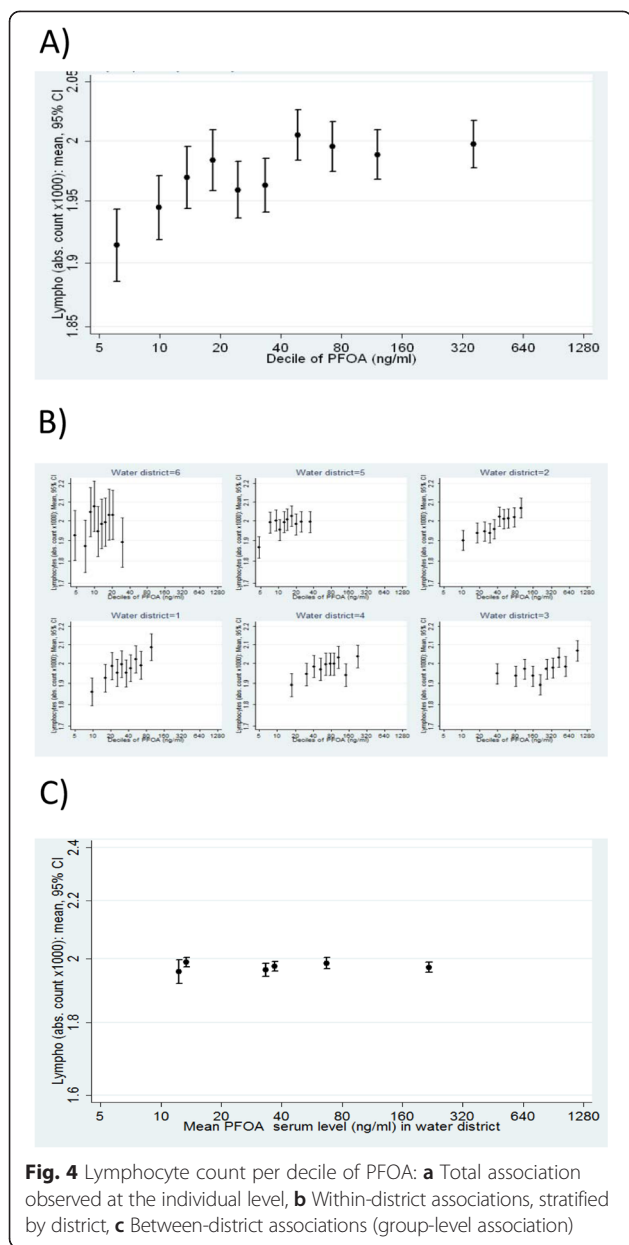


Fig. 4 Lymphocyte count per decile of PFOA: **a** Total association observed at the individual level, **b** Within-district associations, stratified by district, **c** Between-district associations (group-level association)

spurious non-causal exposure-outcome associations. We introduced the basic statistical concepts, discussed the ideas of context effects and cross-level bias, and presented a two-step modelling strategy for practical data analysis within the multilevel framework. From the PFOA study we chose two biomarkers (lymphocyte count and CRP) to illustrate how the approach can be used to improve the robustness of causal claims in cross-sectional analysis; further application examples in the same study are described elsewhere [28–30]. The lymphocyte count showed a strong within-group relation with PFOA but no between-group relation. Thus, we interpreted the observed within-group pattern as a result of individual confounders, i.e. drinking water consumption

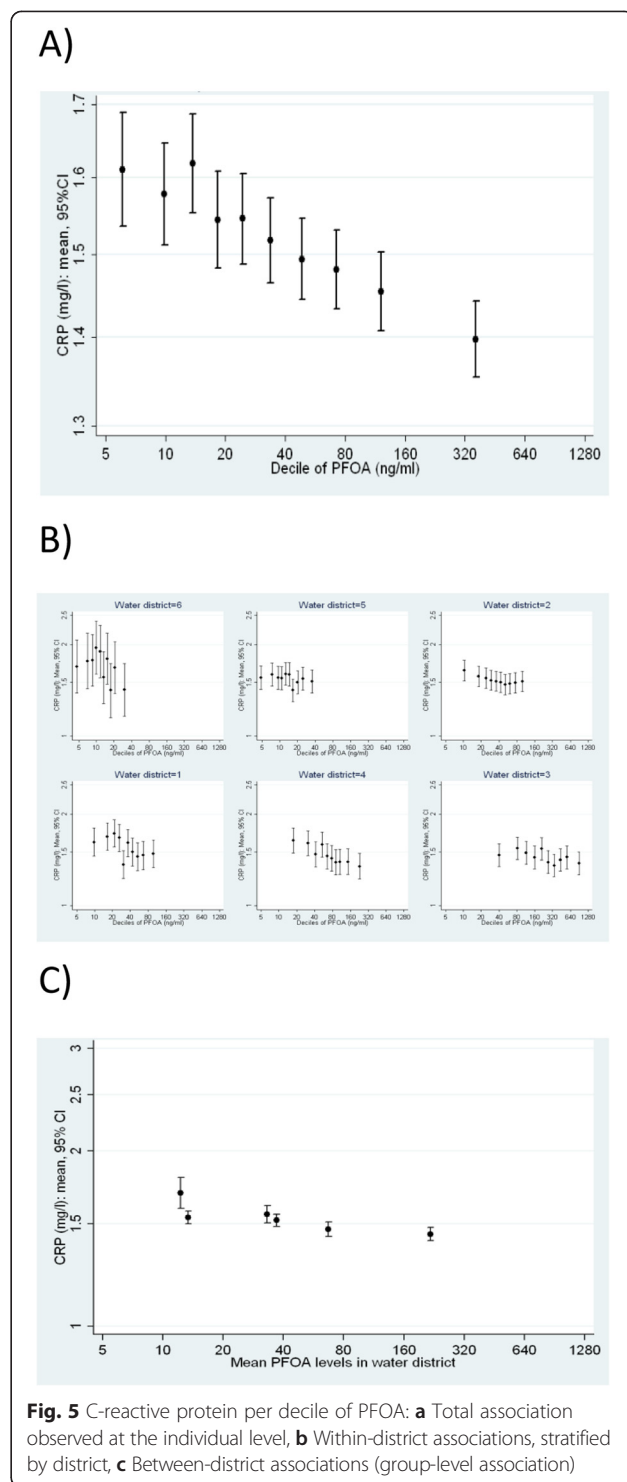


Fig. 5 C-reactive protein per decile of PFOA: **a** Total association observed at the individual level, **b** Within-district associations, stratified by district, **c** Between-district associations (group-level association)

or absorption/excretion of PFOA (e.g. genetic factors), rather than as due to a causal effect of PFOA exposure. In contrast, for CRP we found consistent and significant within- and between-district associations and thus support a causal claim for the effect of PFOA on CRP. In line with this result, we observed a slight clustering of CRP within

districts. This may be interpreted as arising from risk-factor exposure at the group level (e.g. PFOA concentrations in the drinking water assumed to be constant within a particular water district).

The WBGR approach using the district average PFOA serum concentrations as a proxy to estimate the unknown district drinking water concentration at the time of water consumption is related to the statistical framework of instrumental variables [31–33]. An example from the medical literature is the research about the association of CRP and cardiovascular disease. Until genetic Mendelian randomisation studies were performed that introduced genetic factors as instrumental variables to predict biomarkers, it was unclear whether the observed association was causal (i.e. whether CRP causally affects the risk of cardiovascular disease or whether it is merely a marker for a previous cardiovascular). In the present study, district-average PFOA serum concentrations were more proximate to the true exposure (district PFOA water concentration at the time of water consumption) than the individual's PFOA serum concentration. Another application in epidemiology could be the elucidation of the association between unfavourable psychosocial work characteristics and adverse health outcomes, e.g. the association between work-related perception of stress and cardiovascular disease [34]. If aggregated perceived stress perception levels at the department or company level were available, such aggregated value would be more proximate to the unobservable psychosocial adversity of specific work-settings.

The suggested approach is only applicable if there is some clustering of individual exposure concentrations within water districts or units of aggregation ($ICC \gg 0$ and $ICC \ll 1$). The WBGR approach turns a nuisance in straightforward multilevel regression analysis into an advantage for in-depth analysis supporting or refuting causal claims in cross sectional analysis. In the present study, individual PFOA exposure concentrations were substantially clustered within water districts ($ICC = 46\%$, Fig. 3), a not surprising finding since the level of environmental pollution with PFOA was different in the water districts. This high intra-district correlation seemed to be a problem because spatial autocorrelation in exposure may correlate with spatial autocorrelation in disease (e.g. due to spatial clustering in health provision, screening take-up or other risk factors). However, disentangling within- and between-district relations by multilevel modelling helped to reveal spurious, non-causal associations (as demonstrated by the example of biomarker lymphocyte counts). Since a true context effect was unlikely in our environmental example, we interpreted heterogeneous between- and within-district relations as an indicator for estimation bias and non-causality of the observed associations. Individual variation in each

water district may have resulted from variation in daily tap water intake and other factors affecting the bioaccumulation of PFOA. In contrast, the between-group relations obtained by analysing the data at the district level are robust against these individual confounders.

Further applications of WBGR approach in environmental epidemiology are illustrated in the following hypothetical example. Assuming that simple regression analysis shows an association of PFOA with headaches, an individual-level analysis alone does not clarify whether a) PFOA causes (or prevents) headaches or b) "reverse causality" is present, i.e. people with headaches have a different water intake than healthy individuals and thus have different levels of PFOA. In case (a), we would find an association between PFOA and headaches both within and between areas since the higher average PFOA within a given area would increase the prevalence of headaches in this area. In contrast, in case (b) we would find a relation within each district, but between districts there would be little or no correlation as most of the individual PFOA variation is explained by the PFOA concentration in the water supply. Another reason would be the presence of a third (confounding) factor, which was associated with both PFOA and prevalence of headaches. For example, alcohol use may cause headaches and affect water intake. In that case, a naive individual-level analysis might suggest that PFOA prevented headaches, even if the prevalence of headaches was not lower in a district with lower PFOA exposure.

A major limitation of WBGR is that it improves the robustness of causal claims only indirectly by showing up non-causal associations, which are likely due to bias-neglecting confounders and/or effect modifiers on different levels. Homogeneous between- and within- group relations are a necessary but insufficient condition for assessing causal links. Other criteria are needed to further support the causal claim; however, they can often only be assessed by conducting longitudinal studies, for example, a temporal relationship, plausibility, consistency and strength of association. Further details are described in a systematic approach originally elaborated by Hill [5]. A further limitation to the approach is that the average exposure concentrations of geographical areas are not perfect instrument variables as are genetic factors in a Mendelian randomisation study. However, in our application example, detailed environmental experimental and modelling studies [23] substantiated the claim that the average serum PFOA concentrations in a particular water district may serve as a proxy for its PFOA concentrations in the drinking water.

Conclusion

Our methodological work shows that WBGR is an elegant technique for the statistical analysis of clustered

epidemiological data. The statistical approach proposed in this paper may improve the robustness of causal claims of exposure-outcome associations in cross-sectional analysis by unmasking non-causal associations showing up due to hidden confounding. The approach is especially useful for individual-level analysis in environmental epidemiology in which individuals were recruited from different geographical areas with heterogeneous levels of environmental exposure.

Appendix

Mathematical details of within- and between group regression

For a continuous outcome Y and an explanatory variable X (with observations of individual i clustered within groups j) we obtain the multilevel regression model

$$Y_{ij} = \alpha_{00} + \beta_{10} x_{ij} + U_{0j} + R_{ij}. \quad (\text{A1})$$

For simplicity, model (A1) only incorporates one individual-level explanatory variable (X); extensions for more explanatory variables at both individual- and group-levels are straightforward. Model (A1) is called a *mixed-effects model*, incorporating an intercept α_{00} , a *fixed effect* (β_{10}) quantifying the effect of the explanatory variable X and a *random effect* U_{0j} , a random variable which follows a presumed theoretical distribution (usually normal or gamma), and a normal distributed error term R_{ij} . We note that in model (A1) the within- and between group coefficients are forced to be equal, which is an unrealistic assumption for many practical applications. Thus, as an extension, we add the group mean \bar{x}_j as an explanatory variable and obtain the more flexible model

$$Y_{ij} = \alpha_{00} + \beta_{10} x_{ij} + \beta_{01} \bar{x}_j + U_{0j} + R_{ij}, \quad (\text{A2})$$

which allows BGE and WGE to be different.

If we consider model (A2) within a given group, the terms can be reordered as

$$Y_{ij} = (\alpha_{00} + \beta_{01} \bar{x}_j + U_{0j}) + \beta_{10} x_{ij} + R_{ij}, \quad (\text{A3})$$

with the group-specific random intercept ($\alpha_{00} + \beta_{01} \bar{x}_j + U_{0j}$) and the regression coefficient of X within this group (β_{10}). The systematic part is the *within-group regression model*

$$Y = (\alpha_{00} + \beta_{01} \bar{x}_j) + \beta_{10} x \quad (\text{A4})$$

By taking the group average on both sides of equation (A3), we obtain the *between-group regression model*

$$\begin{aligned} \bar{Y}_j &= \alpha_{00} + \beta_{10} \bar{x}_j + \beta_{01} \bar{x}_j + U_{0j} + R_{ij} \\ &= \alpha_{00} + (\beta_{10} + \beta_{01}) \bar{x}_j + U_{0j} + R_{ij}. \end{aligned} \quad (\text{A5})$$

Model (A5) shows that using this parameterization the term $(\beta_{10} + \beta_{01})$ estimates the BGE. If we reject the null hypothesis ($\beta_{01} = 0$) we fit a second model (7) incorporating group-centered explanatory variables, i.e. variables with mean 0 that measure the individuals' deviation within a group from the group mean:

$$Y_{ij} = \tilde{\alpha}_{00} + \tilde{\beta}_{10} (x_{ij} - \bar{x}_j) + \tilde{\beta}_{01} \bar{x}_j + U_{0j} + R_{ij}. \quad (\text{A6})$$

Model (A6) now explicitly estimates BGE ($\tilde{\beta}_{01} = \beta_{10} + \beta_{01}$) and WGE ($\tilde{\beta}_{10} = \beta_{10}$).

Competing interests

All authors disclose any competing interests.

Authors' contribution

BG had the idea for the research and has written the manuscript. The other authors contributed in drafting the manuscript. All authors read and approved the final manuscript.

Acknowledgements

The authors thank the participants of the C8 Health project for their contributions to this study. They also thank Dr Tony Fletcher and Dr Ben Armstrong for their comments on the manuscript and Amy Beierholm and Susan Sills for linguistic corrections. Funding for this work, the *C8 Science Panel Community Study* at London School of Hygiene and Tropical Medicine (LSHTM), comes from the *C8 Class Action Settlement Agreement* (Circuit Court of Wood County, WV, USA) between DuPont and plaintiffs, which resulted from the release of *perfluorooctanoic acid* (PFOA) into drinking water. It is one of the *C8 Science Panel Studies* undertaken by the Court-approved C8 Science Panel established under the same Settlement Agreement. The task of the C8 Science Panel, of which TF is a member, is to undertake research in the Mid-Ohio Valley and subsequently evaluate the results, along with other available information, to determine if there are any probable links between PFOA and disease. Funds were administered by the Garden City Group (Melville, NY), which reports to the Court. This work was further supported by the *Brazilian Conselho Nacional de Desenvolvimento Científico e Tecnológico* [contract no. 400011-2011-0 to BG]. There was no funding by National Institutes of Health (NIH) or, Wellcome Trust. or the Howard Hughes Medical Institute (HHMI).

Author details

¹Mannheim Institute of Public Health, Social and Preventive Medicine, University of Heidelberg, Ludolf-Krehl-Strasse 7-11, Mannheim 68167, Germany. ²Instituto de Saúde Coletiva, Federal University of Bahia, Salvador, Brazil. ³Centro de Pesquisa Gonçalo Muniz, Fundação Oswaldo Cruz (FIOCRUZ), Salvador, Bahia, Brazil.

Received: 21 February 2015 Accepted: 19 June 2015

Published online: 10 July 2015

References

- Morgenstern H. Ecologic Studies. In: Rothman KJ, Greenland S, editors. *Modern Epidemiology*. Lipincott: Philadelphia; 1998. p. 459–80.
- Prentice RL, Sheppard L. Dietary fat and cancer: rejoinder and discussion of research strategies. *Cancer Causes Control*. 1991;2(1):53–8.
- Prentice RL, Sheppard L. Dietary fat and cancer: consistency of the epidemiologic data, and disease prevention that may follow from a practical reduction in fat consumption. *Cancer Causes Control*. 1990;1(1):81–97. discussion 99–109.
- Prentice RL, Pepe M, Self SG. Dietary fat and breast cancer: a quantitative assessment of the epidemiological literature and a discussion of methodological issues. *Cancer Res*. 1989;49(12):3147–56.
- Hill AB. The Environment and Disease: Association or Causation? *Proc R Soc Med*. 1965;58:295–300.

6. Greenland S. Randomization, statistics, and causal inference. *Epidemiology*. 1990;1(6):421–9.
7. Rothman KJ, Greenland S. Causation and causal inference in epidemiology. *Am J Public Health*. 2005;95 Suppl 1:S144–150.
8. Pearl J. Robustness of causal claims. *Proceeding UAI '04 Proceedings of the 20th conference on Uncertainty in artificial intelligence*. Virginia, United States: AUAI Press Arlington; 2004. p. 446–53.
9. Sheppard L. Insights on bias and information in group-level studies. *Biostatistics*. 2003;4(2):265–78.
10. Greenland S. Ecologic versus individual-level sources of bias in ecologic estimates of contextual health effects. *Int J Epidemiol*. 2001;30(6):1343–50.
11. Greenland S, Morgenstern H. Ecological bias, confounding, and effect modification. *Int J Epidemiol*. 1989;18(1):269–74.
12. Rice N, Leyland A. Multilevel models: applications to health data. *J Health Serv Res Policy*. 1996;1(3):154–64.
13. Greenland S. Principles of multilevel modelling. *Int J Epidemiol*. 2000;29(1):158–67.
14. Austin PC, Goel V, van Walraven C. An introduction to multilevel regression models. *Can J Public Health*. 2001;92(2):150–4.
15. Davis JA, Spaeth JL, Huson C. A technique for analyzing the effects of group composition. *Am Sociological Rev*. 1961;26:215–25.
16. Miller KA, Siscovick DS, Sheppard L, Shepherd K, Sullivan JH, Anderson GL, et al. Long-term exposure to air pollution and incidence of cardiovascular events in women. *N Engl J Med*. 2007;356(5):447–58.
17. Snijders T, Bosker R. *Multilevel analysis*. London: Sage Publications; 2000.
18. Hausman JA, Taylor WE. Panel data and unobservable individual effects. *Econometrica*. 1981;49:1377–98.
19. Diez-Roux AV. Multilevel analysis in public health research. *Annu Rev Public Health*. 2000;21:171–92.
20. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika*. 1986;73(1):13–22.
21. Huber PJ, editor. *The behavior of maximum likelihood estimates under nonstandard conditions*. Fifth Berkeley Symposium on Mathematical Statistics and Probability. Berkeley, CA: University of California Press; 1967.
22. White H. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*. 1980;48:817–30.
23. Durkheim E. *Suicide* (translated by Spaulding JA and Simpson G). Glencoe, IL: Free Press; 1951.
24. Frisbee SJ, Brooks Jr AP, Maher A, Flensburg P, Arnold S, Fletcher T, et al. The C8 health project: design, methods, and participants. *Environ Health Perspect*. 2009;117(12):1873–82.
25. Steenland K, Jin C, MacNeil J, Lally C, Ducatman A, Vieira V, et al. Predictors of PFOA levels in a community surrounding a chemical plant. *Environ Health Perspect*. 2009;117(7):1083–8.
26. Steenland K, Tinker S, Frisbee S, Ducatman A, Vaccarino V. Association of perfluorooctanoic acid and perfluorooctane sulfonate with serum lipids among adults living near a chemical plant. *Am J Epidemiol*. 2009;170(10):1268–78.
27. Lopez-Espinosa MJ, Fletcher T, Armstrong B, Genser B, Dhataria K, Mondal D, et al. Association of Perfluorooctanoic Acid (PFOA) and Perfluorooctane Sulfonate (PFOS) with Age of Puberty among Children Living near a Chemical Plant. *Environ Sci Technol*. 2011;45(19):8160–6.
28. Gallo V, Leonardi G, Genser B, Lopez-Espinosa MJ, Frisbee SJ, Karlsson L, et al. Serum Perfluorooctanoate (PFOA) and Perfluorooctane Sulfonate (PFOS) Concentrations and Liver Function Biomarkers in a Population with Elevated PFOA Exposure. *Environ Health Perspect*. 2012;120(5):655–60.
29. Mondal D, Lopez-Espinosa MJ, Armstrong B, Stein CR, Fletcher T. Relationships of Perfluorooctanoate and Perfluorooctane Sulfonate Serum Concentrations Between mother-child pairs in a Population with Perfluorooctanoate Exposure from Drinking Water. *Environ Health Perspect*. 2012;120(5):752–7.
30. Steenland K, Fletcher T, Savitz DA. Epidemiologic evidence on the health effects of perfluorooctanoic acid (PFOA). *Environ Health Perspect*. 2010;118(8):1100–8.
31. Gennettian LA, Magnuson K, Morris PA. From statistical associations to causation: what developmentalists can learn from instrumental variables techniques coupled with experimental data. *Dev Psychol*. 2008;44(2):381–94.
32. Greenland S. An introduction to instrumental variables for epidemiologists. *Int J Epidemiol*. 2000;29(4):722–9.
33. Rassen JA, Brookhart MA, Glynn RJ, Mittleman MA, Schneeweiss S. Instrumental variables I: instrumental variables exploit natural variation in nonexperimental data to estimate causal relationships. *J Clin Epidemiol*. 2009;62(12):1226–32.
34. Kivimaki M, Nyberg ST, Batty GD, Fransson EI, Heikkila K, Alfredsson L, et al. Job strain as a risk factor for coronary heart disease: a collaborative meta-analysis of individual participant data. *Lancet*. 2012;380(9852):1491–7.

Submit your next manuscript to BioMed Central and take full advantage of:

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at
www.biomedcentral.com/submit

