A brief conceptual tutorial of multilevel analysis in social epidemiology: linking the statistical concept of clustering to the idea of contextual phenomenon.

Merlo, Juan; Chaix, Basile; Yang, Min; Lynch, John; Råstam, Lennart

Published in:
Journal of Epidemiology and Community Health

DOI:
10.1136/jech.2004.023473

2005

A brief conceptual tutorial of multilevel analysis in social epidemiology: linking the statistical concept of clustering to the idea of contextual phenomenon

Juan Merlo, Basile Chaix, Min Yang, John Lynch and Lennart Råstam

J. Epidemiol. Community Health 2005;59;443-449
doi:10.1136/jech.2004.023473

THEORY AND METHODS

A brief conceptual tutorial of multilevel analysis in social epidemiology: linking the statistical concept of clustering to the idea of contextual phenomenon

Juan Merlo, Basile Chaix, Min Yang, John Lynch, Lennart Råstam

J Epidemiol Community Health 2005;**59**:443–449. doi: 10.1136/jech.2004.023473

See end of article for authors' affiliations

Correspondence to: Professor J Merlo, Department of Clinical Sciences (Community Medicine), Malmö University Hospital, Faculty of Medicine (Campus Malmö), Lund University, S-205 02 Malmö, Sweden; juan.merlo@med.lu.se

Accepted for publication 19 August 2004

Study objective: This didactical essay is directed to readers disposed to approach multilevel regression analysis (MLRA) in a more conceptual than mathematical way. However, it specifically develops an epidemiological vision on multilevel analysis with particular emphasis on measures of health variation (for example, intraclass correlation). Such measures have been underused in the literature as compared with more traditional measures of association (for example, regression coefficients) in the investigation of contextual determinants of health. A link is provided, which will be comprehensible to epidemiologists, between MLRA and social epidemiological concepts, particularly between the statistical idea of clustering and the concept of contextual phenomenon.

Design and participants: The study uses an example based on hypothetical data on systolic blood pressure (SBP) from 25 000 people living in 39 neighbourhoods. As the focus is on the empty MLRA model, the study does not use any independent variable but focuses mainly on SBP variance between people and between neighbourhoods.

Results: The intraclass correlation (ICC = 0.08) informed of an appreciable clustering of individual SBP within the neighbourhoods, showing that 8% of the total individual differences in SBP occurred at the neighbourhood level and might be attributable to contextual neighbourhood factors or to the different composition of neighbourhoods.

Conclusions: The statistical idea of clustering emerges as appropriate for quantifying ''contextual phenomena'' that is of central relevance in social epidemiology. Both concepts convey that people from the same neighbourhood are more similar to each other than to people from different neighbourhoods with respect to the health outcome variable.

This article has, on the one hand, didactic purposes and is directed to readers disposed to approach multilevel regression analysis (MLRA) in a more conceptual than mathematical way. Readers who wish an alternative or more formal statistical explanation may consult any of the other references on multilevel analysis published elsewhere.[1–5]

On the other hand, and perhaps more important, in this essay we also develop a vision of multilevel analysis[6] that considers measures of health variation[7] (for example, neighbourhood variance, intraclass correlation) for understanding the distribution of health in the general population rather than only applying measures of association (for example, regression coefficients, odds ratios)[8] to understand contextual determinants of individual health. We believe that, so far, measures of health variation have been underused in multilevel epidemiology.

Our aim is to provide a link, which will be comprehensible to epidemiologists, between MLRA techniques and social epidemiological concepts, particularly the analogy between the statistical concept of clustering and the social epidemiological idea of contextual phenomenon.

It is intuitive that people from the same area may be more similar to each other in relation to their health status than to people from other areas. In other words, persons with similar characteristics may have different degrees of health according to whether they live in one area or another because of differing cultural, economic, political, climatic, historical, or geographical contexts.[9] This contextual phenomenon expresses itself as clustering of individual health status within areas. That is, a portion of the health differences among people may be attributable to the areas in which they reside.[6 10]

The notion of contextual phenomenon has a long history in epidemiology and is included under different forms in the Durkheimian concept of social fact,[11] Rose's notion of population disease rates,[12 13] and John Snow's findings on cholera incidence.[14] These three related seminal conceptions are contextual in their nature, and support the idea that knowledge on the distribution and determinants of population health is epistemologically multilevel[15] and needs to consider both people and areas.[10 16 17]

The idea of contextual phenomenon, which could be considered as a core notion in social epidemiology, corresponds to the statistical concept of clustering*—that is, in turn, the main reason for applying multilevel regression techniques. Statistically, it is necessary to use techniques that, like MLRA, consider the dependence of the outcome variable between people from the same area. An important assumption made in usual regression analyses is the independence of individual measures. If this assumption is violated, the results of the regression analysis are biased.[1] However, we have previously emphasised[6] that clustering of individual health within neighbourhoods is not a statistical nuisance that only needs to be considered for obtaining correct statistical estimations, but a key concept in social epidemiology that yields important information by itself.[2 6 19–21] The more the health of the people within a neighbourhood is

Abbreviations: ICC, intraclass correlation; SBP, systolic blood pressure; MLRA, multilevel regression analysis; VPC, variance partition coefficient

alike (as compared with people in other neighbourhoods), the more probable it is that the determinants of individual health are directly related to the contextual environment of the neighbourhood, and/or that social processes of geographical segregation are taking place—that is, similar types of people choose or are forced to reside in a given neighbourhood.

Those aspects are of high significance in social epidemiology as they have value in the context of ideas about the efficacy of focusing intervention to reduce health inequalities on certain geographical areas rather than on specific people only. Measures of variation are important in public health to understand the significance of specific contexts for different individual health outcomes.[7] Traditional measures of association, in contrast with measures of variation, do not inform on the multilevel distribution of health.[6]

Without being a panacea to miraculously fix the ills of a ''risk factor'' epidemiology that seems inappropriate for assessing the impact of the context,[23] MLRA is a suitable statistical technique that can be used to operationalise conceptual schemas in multilevel epidemiology.

In this essay we explain how to investigate whether a given health phenomenon (for example, systolic blood pressure) has a contextual dimension. Using this research question, we introduce the ''empty'' MLRA model. This model is the simplest form of MLRA as it does not include any covariate but focuses only on how health differences are distributed between people and between areas. Along the explanation of the empty model, we present figures that permit a visual comprehension of MLRA concepts such as residuals, partitioning of variance at different levels, and the idea of clustering and intraclass correlation.

THE ''EMPTY'' MLRA MODEL

To explain MLRA we use an example based on hypothetical data. The population of the example consists of 25 000 subjects, 35 to 64 years old, living in the 39 neighbourhoods of an imaginary city. The individual outcome variable is systolic blood pressure (SBP), and we assume that it is continuous and follows a normal distribution. As this article explains the empty MLRA model, we do not use any independent variable but focus only on the mean and variance of SBP.

The example was adapted from a real empirical investigation that analysed countries rather than neighbourhoods.[10] This essay is based on simulated data and, therefore, the results presented in this article should not be used as empirical evidence. For all analysis, we use the software MLwiN version 1.1 developed by Goldstein's research group†.[24]

The reason for naming this model ''empty'' is that it does not include any explanatory variables but only estimates the city SBP mean and the neighbourhood and individual differences in SBP on the basis of the study sample. We present below a very simple equation of the model that will be clear to readers not trained to read formal statistical notations. Readers who wish a formal statistical explanation are referred elsewhere.[2 25 26]

$SBP_I = SBP_C + E_{N-c} + E_{I-c}$

SBP_I = SBP of an individual in a neighbourhood

SBP_C = Mean SBP of the city

E_{N-c} = Difference between the city SBP mean and the neighbourhood SBP mean (also known as neighbourhood ''shrunken residual'')

E_{I-c} = Difference between the neighbourhood SBP mean and the individual SBP value (also known as ''individual residual'')

In MLRA both people and neighbourhoods are assumed to be randomly sampled from a population of persons and a population of neighbourhoods. It is assumed that the residuals are normally distributed and that there is independence between the individual residuals and the neighbourhood residuals. MLRA presents advantages compared with the common analysis of variance.

The model presented above shows that the SBP value of a person living in a neighbourhood (SBP_I) is equal to the mean SBP in the city (SBP_C) plus the predicted neighbourhood difference from the city mean (that is, neighbourhood shrunken residual [E_{N-c}]) plus the individual difference from the neighbourhood mean (that is, individual residual [E_{I-c}]).

Partitioning overall differences in SBP

The presence of neighbourhood and individual residuals in the empty multilevel model just shows that SBP varies both at the individual and at the neighbourhood level. The main intent of the empty model is to partition the total variance in SBP in the city (V_{Total}) into a variance that occurs between neighbourhoods (V_N) and a variance that occurs between people (V_I) as shown in the equation 1, illustrated in figure 1, and calculated in table 1.

$$V_{Total} = V_N + V_I \qquad (1)$$

In figure 1 we can visualise the empty model and the concepts that it conveys. In this figure the neighbourhood differences from the city mean represent the shrunken residuals. The figure shows that multilevel structures convey information on variability both between and within neighbourhoods. The variance is a summary of the differences. The higher the variance, the larger the differences are. In figure 1 the brackets show that the total variance is the sum of the between neighbourhood variance and the within neighbourhood variance.

Figure 2 shows the individual and neighbourhood SBP values used in our example. We can see that each neighbourhood has a specific SBP mean (black dots) that differs from the city mean (130 mm Hg) by a certain amount of mm Hg. This difference is the neighbourhood level raw residual.

Single level individual studies compared with MLRA

In table 1 we can see that the empty model gives evidence of both between individual ($V_I = 433.4$) and between neighbourhood ($V_N = 36.2$) variance in SBP. If we combine the variance from both levels to give a total variance, we see that this total variance ($V_{Total} = 468.1$) is similar to the variance

*According to the ideas of Durkheim (1858–1917) people belonging to a specific community share a collective conscience (common social values and norms that are formed by human relations and interactions and that generate collective feelings of solidarity and connectedness). This collective conscience operates creating what Durkheim called ''social cohesion'' to bind the social structure together. Understood in this way, the social group emerges as an independent social fact rising over and above individual circumstances, and going beyond the sum of the people that compose it.[11] Thus collective characteristics shape the health of the population in a way that cannot be reduced to individual characteristics. A classic example concerns population differences in suicide rates. Even if within each area the people at risk of committing suicide are not the same in different time periods, the differences between populations in suicide rates are fairly stable over time. This fact suggests the existence of a contextual phenomenon that conditions a clustering of individual suicide risk within areas. In other words, some part of the total differences in health between people might be as a consequence of the differences between the areas where the people live. Analogous consideration can be made when interpreting John Snow's findings on cholera incidence in different areas of London[14] and the ideas of Geoffrey Rose on sick people and sick populations.[10 12 18]

†For a review on other programs suitable for MLRA see the Centre for Multilevel Modelling, Institute of Education, London (http://multilevel.ioe.ac.uk/ softrev/index.html).

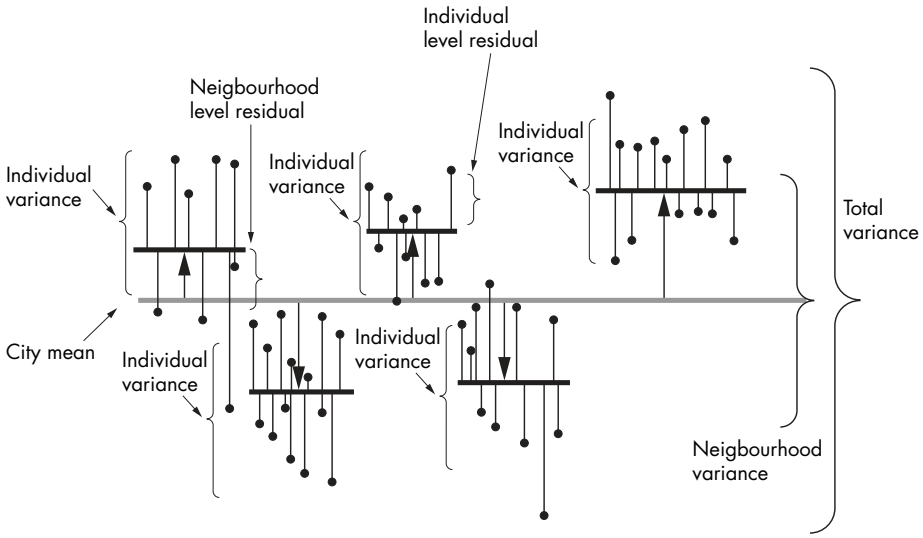

Figure 1 Multilevel information. In this figure the neighbourhood residuals are represented by the length of the pillows between the city SBP mean, represented by a grey colour, and the neighbourhood SBP means represented by thick black horizontal lines. The individual residuals are represented by the length of the vertical lines between the neighbourhood means and the individual SBP values represented by black circles at the top of thin lines. In this figure we do not have any explanatory variable (that is, this figure corresponds to an ''empty'' model) as we are only interested in analysing how individual blood pressure differences are partitioned in a variability that exists between people from the same neighbourhood and a variability that exists between neighbourhoods. In this figure we can imagine that the neighbourhood means (short thick lines) pull up or pull down all the individual SBP values belonging to the same neighbourhood, even if individual level variability remains within neighbourhoods. The mathematical expression of the intraclass correlation can be visually understood in figure 1. Figure 1 is a graphic combination of figures 3 and 4.

obtained by a simple individual level analysis using descriptive statistics. Reading table 1, you can understand intuitively that a portion of the individual level variance is in fact neighbourhood level variance.

Imagine that figure 3 represents the distribution of individual SBP in the population of the city described in the example. As in individual single level analysis we have individual data only, the fact that people are grouped within neighbourhoods is neglected. In figure 3 we merely see differences between the individual SBP values and the mean SBP value of the whole city (the single level individual residuals). We are unable to distinguish the differences between the mean blood pressure in each neighbourhood and the overall mean blood pressure in the city. In single level individual based designs, we tend to neglect possible neighbourhood effects. This oversimplified approach has been termed the individualist fallacy.[27]

Single level ecological studies compared with MLRA

Table 1 also shows the between neighbourhood variance obtained by an ecological analysis performed by aggregation of the individual SBP values at the neighbourhood level. In the ecological analysis we estimate the mean SBP for each neighbourhood from the sample of people in each neighbourhood, and then we compute the variance of these estimated means. The ecological variance computed in this way overstates the neighbourhood variance because it also includes variation attributable to sampling error (imprecision) in the estimates of each neighbourhood mean SBP.

Table 1 Multilevel, individual and ecological linear regression analysis of systolic blood pressure (SBP) in 25 000 people living in the 39 neighbourhoods (hypothetical data)

	Multilevel regression analysis (empty model)	Single level individual analysis*	Single level ecological analysis†
Fixed effects			
Mean SBP of the city	130.2	130.2	130.0
Random effects			
Multilevel measures of health variation			
Variance (SE)			
Between neighbourhoods	36.2	–	37.0
Between individuals	433.4	468.1	
Intraclass correlation (ICC)	ICC = 0.08	–	–
Deviance (goodness of fit of the model)‡	222764	224532	–

''Fixed effects'' and ''random effects'' are expressions that are often used in MLRA. In very simple terms, fixed effects are used to model means whereas random effects are used to model variance. *In the individual level analysis we ignore the existence of the neighbourhood level. †In the single level ecological analysis we aggregate the individual level information on SBP to the neighbourhood level, neglecting the existence of the individual level. ‡The deviance expresses the goodness of fit of the model, and was calculated using the −2×log likelihood (IGLS). The Bayesian deviance information criterion (DIC) gave similar results. The multilevel models fit the data much better than the single level individual analysis. The ecological analysis is not comparable as it is based on 39 neighbourhoods rather than on 25 000 people. We refer elsewhere for an extended explanation of the deviance statistic.[44 45]

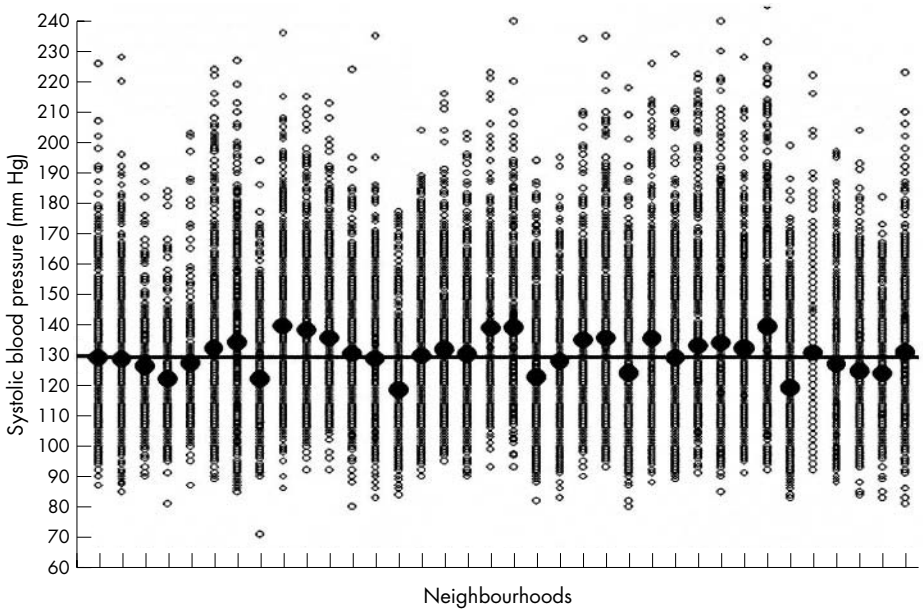

Figure 2 The figure shows the actual SBP values used in our example. The large black dots represent the neighbourhood means. The small circles represent the individual SBP values within neighbourhoods. The horizontal black line represents the city SBP mean.

The ecological variance is rather similar to the between neighbourhood variance in SBP obtained by the MLRA. It is patently clear that the single level ecological analysis neglects the existence of individual level variance. Figure 4 illustrates that in an ecological analysis we are unable to observe differences between people (variation in blood pressure within a neighbourhood), but we can distinguish differences between the mean blood pressure of each neighbourhood and the mean blood pressure of the whole city (that is, the neighbourhood residuals of an ecological analysis).

In single level ecological analysis, we consider all information as if it were at the neighbourhood level, neglecting possible individual components (this oversimplified approach has been termed the sociological fallacy).[27]

Today it is well known that neither single level individual nor ecological analyses are suitable for effectively investigating contextual effects.[22]

The intraclass correlation (ICC) or variance partition coefficient (VPC)

It is seen in figure 1 that all people living in the same neighbourhood share a common level of blood pressure that differs from the city mean in an amount that corresponds to the neighbourhood residual. Therefore, we often speak about "differences between neighbourhoods" and "differences between people within neighbourhoods". Together the individual and the neighbourhood variance components represent the total differences in SBP. We can see that a portion of the total individual SBP difference is at the neighbourhood level, and in the empty model we can quantify this aspect by computing ICC. As equation 2 shows, the multilevel ICC is the proportion of the variance in SBP that occurs at the neighbourhood level. In this sense the ICC is a variance partition coefficient (VPC).[1]

It can be seen in figures 3 and 4 that in single level analysis we are unable to calculate the ICC, because information on how variance is partitioned at different levels is not available.

The ICC equation is intuitive and can also be understood by observing figure 1.

$$\mathrm{ICC} = \frac{V_N}{V_N + V_I} \qquad (2)$$

In this formula V_I is the variance between people from the same neighbourhood (1st level variance) and V_N is the variance between neighbourhoods (2nd level variance).

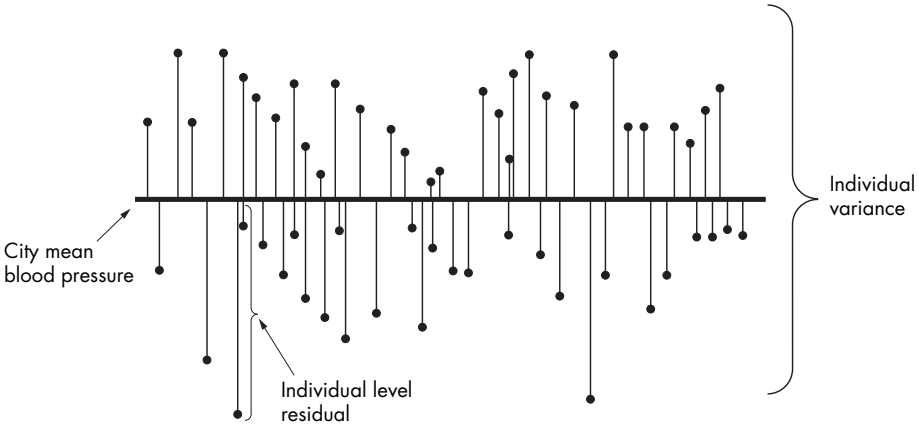

Figure 3 Single level individual information. This figure represents the distribution of individual SBP in the population of the city when we have only single level individual based information. The fact that people are grouped within neighbourhoods is neglected, as we only have individual level data. In this figure the length of the thin vertical line from the black spot to the thick horizontal line represents the individual differences in blood pressure compared with whole city mean (the individual level residuals). The individual variance in single level individual studies is an average summary of these differences. In single level individual analysis we consider all information as if it were at the individual level neglecting possible neighbourhood components.

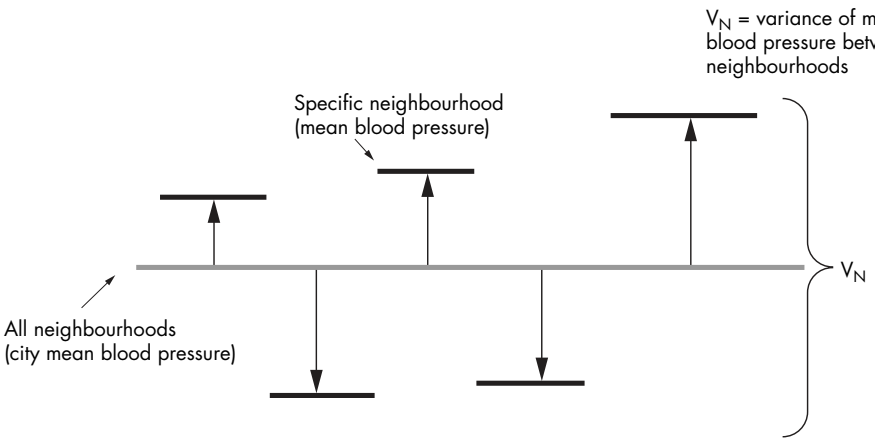

Figure 4 Single level ecological information. In this figure all individual SBP values are aggregated at the neighbourhood level to obtain the neighbourhood mean. We can distinguish differences between the mean blood pressure of each neighbourhood and the mean blood pressure of the whole city (the neighbourhood residuals). These residuals are represented by thick black horizontal lines at the top of a pillow. The neighbourhood variance is a summary of the differences between neighbourhoods. We are unable to observe differences between people (variation in blood pressure within neighbourhoods). In single level ecological analysis we consider all information as if it were at the neighbourhood level neglecting individual components.

As variance can only be positive, according to equation 2 the ICC is necessarily between 0 and 1. Table 1 shows that the ICC, which measures individual SBP clustering at the neighbourhood level, is equal to 0.08. Therefore, in our example 8% of the total individual differences in SBP are at the neighbourhood level. On these grounds, we might conclude that there is some evidence for a possible neighbourhood contextual phenomenon shaping a common individual SBP level. Alternatively, this clustering might be attributable to the different composition of neighbourhoods.

The name ''correlation'' suggests that the ICC expresses the similarity in health status (in our example SBP) between two persons in the same neighbourhood. An ICC equal to 1 would inform us that all the people in a neighbourhood have an identical SBP level (that is, 100% of the total individual differences are at the neighbourhood level), and an ICC equal to 0 that the people do not share any neighbourhood related common level of SBP.

A high ICC value informs us that neighbourhoods are very important in understanding individual differences in health. On the other hand, an ICC of 0 would suggest that the neighbourhoods are similar to random samples taken from the city and suggest that neighbourhoods are not relevant to understanding SBP differences. Snijders also gives a didactic example of this concept (page 18).[2] When the ICC is 0, the suitability of performing a multilevel analysis is less obvious. In the absence of a multilevel structure, a single level individual analysis is appropriate.

We may be interested in knowing if the ICC is statistically different from 0. The simplest method would be to perform a statistical test of the neighbourhood variance.[1–3] When the neighbourhood level variance is not significant, there is no justification for computing the ICC. However, when testing the neighbourhood variance you need to consider the statistical power in MLRA considering that it depends more on the number of neighbourhoods than on the number of people.[2] Remember that absence of evidence is not evidence of absence.[28]

If the ICC is 0, it does not necessarily mean that the neighbourhood context is not important compared with individuals' factors. Rather, an alternative reason could be that the geographical boundaries we use to define the actual neighbourhoods do not correspond with the boundaries that shape the relevant environment for the concrete individual health outcome. An ICC close to 0 in an empty model may hide considerable neighbourhood variability that would only appear in more complex models. Moreover, a small ICC does not prevent the existence of significant associations between neighbourhood variables and individual health as comparatively small variance between neighbourhood means may give enough contrast of exposure to detect associations.[6 25] These aspects are more extensively discussed in companion papers.[29 30]

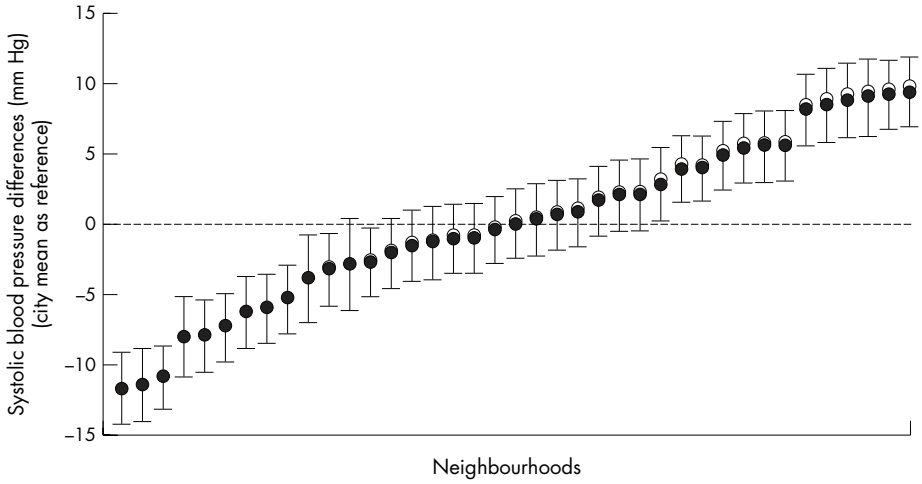

Figure 5 Here neighbourhoods are ranked according to the mean SBP using the whole city mean SBP as reference in the comparisons. The neighbourhood values are the ''shrunken residuals'' (black circles) and the raw residuals (white circles). We provide 95% confidence intervals obtained in the multilevel regression analysis.

What this paper adds

- We provide a link—comprehensible to epidemiologists—between multilevel regression techniques and social epidemiological concepts, particularly the analogy between the statistical concept of clustering and the social epidemiological idea of contextual phenomenon.
- We develop a vision of multilevel analysis that considers measures of health variation (for example, neighbourhood variance, intraclass correlation) for understanding the distribution of health in the general population rather than only applying measures of association (for example, regression coefficients, odds ratios) to understand contextual determinants of individual health.
- Measures of health variation have been underused in multilevel epidemiology.
- Statistical measures of clustering emerge as appropriate for quantifying ''contextual phenomena'', which is of central relevance in social epidemiology.

Policy implications

- It is important that political decisions are grounded in appropriate analysis. This study explains a modern methodology of analysis that can be applied in this context.
- Multilevel analyses can be used to identify the relevance of the neighbourhood or other societal boundaries for understanding health inequalities.
- Our study has value in the context of ideas about the efficacy of focusing intervention to reduce health inequalities on certain geographical boundaries rather than on people only.
- Politicians should always consider the fact that the health of the citizens may depend on their context, which deserves to be investigated and accounted for when planning public health interventions.

''Shrunken'' neighbourhood level residuals

An extra comment on how neighbourhood level residuals are calculated in MLRA is necessary as these residuals are often used in epidemiology and community health studies to rank second level units (for example, hospitals) and investigate geographical differences in health.[31–33]

In the simplest case, the raw residual is the difference between the city and the neighbourhood mean SBP. The shrunken neighbourhood residual (E_{N-c}) obtained with the multilevel regression is then calculated a posteriori by multiplying the raw neighbourhood residual by a shrinkage factor (SF) shown in equation 3:

$$SF = \frac{V_N}{V_N + \frac{V_I}{N_n}} \quad (3)$$

Obviously SF has a value between 0 and 1. The neighbourhood ''shrunken'' residual is calculated by weighting the raw residual with SF as in equation 4:

$$E_{N-c} = \text{Neighbourhood ''raw residual''} \times SF \quad (4)$$

The neighbourhood ''shrunken residuals'' are computed using the raw residuals, the estimated variances, and the number of people in the neighbourhood (N_n). MLRA can be performed even when the number of people (1st level units) within each neighbourhood (2nd level unit) is very different. The fewer the number of people in a neighbourhood, or the higher the variability within neighbourhoods as compared with the variability between neighbourhoods, the more important the shrinkage and the more the value of the neighbourhood residual will be shrunken towards 0. The value (SBP_c+E_{N-c}) is also termed ''posterior mean''.[1]

Computing these shrunken residuals may be viewed as disentangling the proportion of each residual that may be attributed to true variations between neighbourhoods from that proportion that might better be attributed to random variations.[34] Rather than only considering the neighbourhood level variance as a summary of the variations that exist between neighbourhoods, the shrunken residuals inform how each specific neighbourhood differs from the city mean.

In figure 5 we have ranked the neighbourhoods according to their shrunken residual as explained above. The raw residuals are represented by white circles and in our example are very similar to the shrunken residuals due to the high number of individuals in each neighbourhood. The bars around each neighbourhood residual represent the 95% confidence intervals. It can be concluded that in these hypothetical data many of the neighbourhoods present a SBP that differs from the city SBP mean (represented by a dotted line).

CONCLUSIONS

We have shown in a basic way that the simple investigation of how differences in SBP are partitioned between the individual level and the neighbourhood level provides relevant epidemiological information. Both the statistical idea of clustering (that is, ICC) and the social epidemiological concept of contextual phenomenon convey that people from the same neighbourhood are more similar to each other than to people from different neighbourhoods with respect to the health outcome variable. For this reason epidemiological measures of clustering as the ICC emerge as appropriate for identifying and quantifying ''contextual phenomena'',[6 35] which is of central relevance in social epidemiology.[6 7 36]

In companion articles we explain more complex MLRA models that include individual and neighbourhood level variables.[29 37] In these articles we illustrate that the importance of the context for understanding health differences may differ for people with different characteristics. We also clarify that contextual factors may modify the effect of individual characteristics on health, and that individual and contextual factors can be used to explain compositional and contextual neighbourhood differences in health. We explain that measures of association between contextual characteristics and individual health, being important for understanding multilevel causal pathways, do not allow for assessing the multilevel distribution of health outcomes.

Studying multilevel health variation presents comparatively few complications and yields measures that are intuitively easy to understand when the outcome of interest meets the conditions for linear regression analysis. However, when the outcome is not continuous, interpreting measures of variation is less easy and it is the subject of this investigation.[38 39] Appropriate measures are, however, already available,[20 35 40–43] and we explain these measures in a

companion paper. In any case, most epidemiological concepts that can be operationalised by multilevel linear regression analysis are of general validity and can be applied to any type of health outcomes.

Our essay may help to provide more insight into the use of measures of health variation based on the random effects of the multilevel models, and emphasise the decisive part they should play in social epidemiology and community health research. Statistical measures of clustering emerge as appropriate for quantifying ''contextual phenomena'', which is of central relevance in social epidemiology.

ACKNOWLEDGEMENTS

We thank Klaus Larsen for his comments on the manuscript and Beatriz Gonzalez Lopez-Valcarcel and other anonymous referees for their constructive critics.

Authors' affiliations

J Merlo, L Råstam, Department of Clinical Sciences (Community Medicine), Malmö University Hospital, Lund University, Malmö, Sweden
B Chaix, Research Team on the Social Determinants of Health and Healthcare, National Institute of Health and Medical Research, Paris, France
M Yang, Institute of Community Health Sciences, Queen Mary University of London, London, UK
J Lynch, Department of Epidemiology, Center for Social Epidemiology and Population Health, University of Michigan, Ann Arbor, USA

Funding: this study is supported by grants from FAS (Swedish Council for Working Life and Social Research) for the projects ''Development and application of multilevel analysis in pharmacoepidemiology and social medicine'' (principal investigator: Juan Merlo, number 2002-054) and ''Socioeconomic disparities in cardiovascular diseases - a longitudinal multilevel analysis'' (principal investigator: Juan Merlo, number 2003-0580).

Competing interests: none declared.

REFERENCES

1 **Goldstein H**. *Multilevel statistical models*. 3rd ed. London: Hodder Arnold, 2003.
2 **Snijders TAB**, Bosker RJ. *Multilevel analysis: an introduction to basic and advanced multilevel modeling*. 1st ed. Thousand Oaks, CA: Sage, 1999.
3 **Raudenbush SW**, Bryk AS. *Hierarchical linear models: applications and data analysis methods*. 2nd ed. Newbury Park, CA: Sage, 2002.
4 **Duncan C**, Jones K, Moon G. Context, composition and heterogeneity: using multilevel models in health research. *Soc Sci Med* 1998;**46**:97–117.
5 **Subramanian SV**, Jones K. Multilevel methods for public health research. In: Kawachi I, Berkman LF, eds. *Neighbourhoods and health*. Oxford: Oxford University Press, 2003.
6 **Merlo J**. Multilevel analytical approaches in social epidemiology: measures of health variation compared with traditional measures of association. *J Epidemiol Community Health* 2003;**57**:550–52.
7 **Boyle MH**, Willms JD. Place effects for areas defined by administrative boundaries. *Am J Epidemiol* 1999;**149**:577–85.
8 **Diez Roux AV**, Merkin SS, Arnett D, *et al*. Neighborhood of residence and incidence of coronary heart disease. *N Engl J Med* 2001;**345**:99–106.
9 **Macintyre S**, Elleway A. Ecological approaches: rediscovering the role of the physical and social environment. In: Berkman LF, Kawachi I, eds. *Social epidemiology*. New York: Oxford University Press, 2000:332–48.
10 **Merlo J**, Asplund K, Lynch J, *et al*. Population effects on individual systolic blood pressure: a multilevel analysis of the World Health Organisation MONICA Project. *Am J Epidemiol* 2004;**159**:1168–79.
11 **Durkheim E**. *The rules of sociological method*. 8th ed. New York: Free Press of Glencoe, 1964.
12 **Rose GA**. *The strategy of preventive medicine*. Oxford: Oxford University Press, 1992:138.
13 **Schwartz S**, Diez-Roux AV. Commentary: causes of incidence and causes of cases—a Durkheimian perspective on Rose. *Int J Epidemiol* 2001;**30**:435–9.
14 **Snow J**. *Snow on cholera*. (A reprint of two papers by John Snow, MD, together with a biographical memoir by BW Richardson, MD, and an introduction by Wade Hampton Frost, MD). New York: The Commonwealth Fund, 1936.
15 **Diez-Roux AV**. Multilevel analysis in public health research. *Annu Rev Public Health* 2000;**21**:171–92.
16 **Koopman JS**, Lynch JW. Individual causal models and population system models in epidemiology. *Am J Public Health* 1999;**89**:1170–4.
17 **Kaplan GA**. What is the role of the social environment in understanding inequalities in health? *Ann N Y Acad Sci* 1999;**896**:116–19.
18 **Rose GA**. Sick individuals and sick populations. *Int J Epidemiol* 2001;**3**:427–32.
19 **Merlo J**, Ostergren PO, Hagberg O, *et al*. Diastolic blood pressure and area of residence: multilevel versus ecological analysis of social inequity. *J Epidemiol Community Health* 2001;**55**:791–8.
20 **Petronis KR**, Anthony JC. A different kind of contextual effect: geographical clustering of cocaine incidence in the USA. *J Epidemiol Community Health* 2003;**57**:893–900.
21 **Merlo J**, Lynch JW, Yang M, *et al*. Effect of neighborhood social participation on individual use of hormone replacement therapy and antihypertensive medication: a multilevel analysis. *Am J Epidemiol* 2003;**157**:774–83.
22 **Morgenstern H**. Ecological studies. In: Rothman KJ, Greenland S, eds. *Modern epidemiology*. Philadelphia: Lippincott-Raven, 1998:459–80.
23 **Krieger N**. A glossary for social epidemiology. *Epidemiol Bull* 2002;**23**:7–11.
24 **Rasbash J**, Browne W, Goldstein H, *et al*. *A user's guide to MLwiN*. Version 2.1st ed. London: Multilevels Models Project, Institute of Education, University of London, 2000.
25 **Bingenheimer JB**, Raudenbush SW. Statistical and substantive inferences in public health: issues in the application of multilevel models. *Annu Rev Public Health* 2004;**25**:53–77.
26 **Rasbash J**, Steele F, Browne W. *A user's cuide to MLwiN*. Version 2.0. Version 2.1st ed. London: Centre for Multilevel Modelling Institute of Education University of London, 2003.
27 **Diez Roux AV**. A glossary for multilevel analysis. *J Epidemiol Community Health* 2002;**56**:588–94.
28 **Altman DG**, Bland JM. Absence of evidence is not evidence of absence. *BMJ* 1995;**311**:485.
29 **Merlo J**, Yang M, Chaix B, *et al*. A brief conceptual tutorial of multilevel analysis in social epidemiology. Investigating contextual phenomena in different groups of people. *J Epidemiol Community Health* (in press).
30 **Merlo J**, Yang M, Chaix B, *et al*. A brief conceptual tutorial of multilevel analysis in social epidemiology. Interpreting neighbourhood differences and the effect of neighbourhood characteristics on individual health. *J Epidemiol Community Health* (in press).
31 **Leyland AH**, Boddy FA. League tables and acute myocardial infarction. *Lancet* 1998;**351**:555–8.
32 **Goldstein H**, Spiegelhalter D. League tables and their limitations: statistical issues in comparisons of institutional performance. *Journal of the Royal Statistical Society Soc A* 1996;**159**:385–443.
33 **Merlo J**, Ostergren PO, Broms K, *et al*. Survival after initial hospitalisation for heart failure: a multilevel analysis of patients in Swedish acute care hospitals. *J Epidemiol Community Health* 2001;**55**:323–9.
34 **Burton P**, Gurrin L, PS. Extending the simple linear regression model to account for correlated responses: an introduction to generalized estimating equations and multi-level mixed modelling. *Stat Med* 1998;**17**:1261–91.
35 **Chaix B**, Merlo J, Bobashev G, *et al*. Re: ''Detecting patterns of occupational illness clustering with alternating logistic regressions applied to longitudinal data''. *Am J Epidemiol* 2004;**160**:505–6.
36 **Rodriguez G**, Goldman N. An assessment of estimation procedures for multilevel models with binary responses. *Journal of the Royal Statistical Society Soc A* 1995;**158**:73–8.
37 **Merlo J**, Chaix B, Yang M, *et al*. A brief conceptual tutorial of multilevel analysis in social epidemiology: linking the statistical concept of clustering to the idea of contextual phenomenon. *J Epididemiol Community Health* (in press).
38 **Browne W**, Subramanian S, Jones K, *et al*. Variance partitioning in multilevel logistic models that exhibit over-dispersion. (http://multilevel.ioe.ac.uk/team/materials/vpmlmeo.pdf)
39 **Goldstein H**, Browne W, Rasbash J. Partitioning variation in generalised linear multilevel models. *Understanding Statistics* 2002;**1**:223–32.
40 **Petronis KR**, Anthony JC, Merlo J. Social epidemiology, intra-neighbourhood correlation, and generalised estimating equations. *J Epidemiol Community Health* 2003;**57**:914.
41 **Petronis KR**, Anthony JC. Perceived risk of cocaine use and experience with cocaine: do they cluster within US neighborhoods and cities? *Drug Alcohol Depend* 2000;**57**:183–92.
42 **Larsen K**, Petersen JH, Budtz-Jorgensen E, *et al*. Interpreting parameters in the logistic regression model with random effects. *Biometrics* 2000;**56**:909–14.
43 **Larsen K**, Merlo J. Appropriate assessment of neighborhood effects on individual health -integrating random and fixed effects in multilevel logistic regression. *Am J Epidemiol* 2005;**161**:81–8.
44 **Snijders TAB**, Bosker RJ. *Testing a model specification, in multilevel analysis—an introduction to basic and advanced multilevel modeling*. Thousand Oaks, CA: Sage, 1999:86–98.
45 **Browne WJ**. *MCMC estimation in MLwiN*. Version 2.0. London: Centre for Multilevel Modelling, Institute of Education. University of London, 2003:297.