A joint modeling of neighborhood effects on participation in the RECORD Cohort Study and neighborhood effects on type 2 diabetes: bias assessment and correction

Basile Chaix,[a,b] *Nathalie Billaudeau,*[a,b] *Frédérique Thomas,*[c] *Sabrina Havard,*[a,b] *David Evans,*[a,b,d] *Yan Kestens,*[e,f] *Kathy Bean*[c]

[a]Inserm, U707, Research Unit in Epidemiology, Information Systems, and Modeling, Paris, France

[b]Université Pierre et Marie Curie-Paris6, UMR-S 707, Paris, France

[c]Centre d'Investigations Préventives et Cliniques, Paris, France

[d]EHESP School of Public Health, Rennes, France

[e]Centre de Recherche du Centre Hospitalier de l'Université de Montréal, Montreal, Canada

[f]Social and Preventive Medicine Department, Université de Montréal, Montreal, Canada

Type of manuscript: Original Article

Running head: Neighborhood-related selective participation

Correspondence:

Basile Chaix

Inserm U707, Faculté de Médecine Saint-Antoine,

27 rue Chaligny, 75012, Paris, France

Tel.: + 33 1 44 73 86 64

Email: chaix@u707.jussieu.fr

Sources of financial support:

As part of the RECORD project, the present work is funded by the National Research Agency (Agence Nationale de la Recherche) (Health–Environment Program 2005 #00153 05); the Institute for Public Health Research (Institut de Recherche en Santé Publique); the National Institute for Prevention and Health Education (Institut National de Prévention et d'Education pour la Santé) (Prevention Program 2007 074/07-DAS); the National Institute of Public Health Surveillance (Institut de Veille Sanitaire) (Territory and Health Program); the French Ministries of Research and Health (Epidemiologic Cohorts Grant 2008); the National Health Insurance Office for Salaried Workers (Caisse Nationale d'Assurance Maladie des Travailleurs Salariés); the Ile-de-France Health and Social Affairs Regional Direction (Direction Régionale des Affaires Sanitaires et Sociales d'Île-de-France); the Ile-de-France Public Health Regional Group (Groupement Régional de Santé Publique); the City of Paris (Ville de Paris); and the Ile-de-France Youth and Sports Regional Direction (Direction Régionale de la Jeunesse et des Sports).

Acknowledgements:

We would like to express our gratitude to the institutions that provided financial support for the RECORD project. We particularly thank Alfred Spira, head of the French Institute for Public Health Research, for his advice and support. We are also grateful to Danièle Mischlich from the Ile-de-France Health and Social Affairs Regional Direction for her support in our project. We are grateful to Insee, the French National Institute of Statistics and Economic Studies, which provided support for the geocoding of the RECORD participants and allowed us to access to

relevant geographical data (with special thanks to Aline Désesquelles, Pascale Breuil, and Jean-Luc Lipatz). We thank Geoconcept for allowing us to access to the Universal Geocoder software. Regarding the geographical data used in the present analysis, we are also grateful to Paris-Notaires, the National Geographic Institute, the Institute of Planning and Urbanism from the Paris Region, and the Authority for Public Transport in the Paris Region. We also thank the Caisse Nationale d'Assurance Maladie des Travailleurs Salariés (CNAM-TS, France) and the Caisse Primaire d'Assurance Maladie de Paris (CPAM-P, France) for helping make this study possible.

A joint modeling of neighborhood effects on participation in the RECORD Cohort Study and neighborhood effects on type 2 diabetes: bias assessment and correction

ABSTRACT

Background: Studies of neighborhood effects on health based on cohort data are subject to bias induced by neighborhood-related selective study participation.

Methods: Using the RECORD Cohort Study (REsidential Environment and CORonary heart Disease, n = 7233, Paris metropolitan area, France), we performed a separate and joint modeling of neighborhood determinants of study participation and type-2 diabetes. We sought to identify neighborhood-related selective participation processes and account for their biasing effect on the associations with diabetes.

Results: After controlling for individual characteristics, study participation was higher for people residing close to the health centers and in neighborhoods with a high income, high dwelling values, high proportion of the population looking for work, and a low built surface and building height (contextual effects adjusted for each other). After individual-level adjustment, the prevalence of diabetes was elevated in neighborhoods with lowest levels of educational attainment (prevalence odds ratio = 1.56, 95% credible interval: 1.06, 2.31). Neighborhood effects on participation did not bias the association between neighborhood education and diabetes. However, residual geographic variations in participation weakly biased the neighborhood education–diabetes association. Bias correction through the joint modeling of neighborhood determinants of participation and diabetes resulted in a 18% decrease of the log prevalence odds ratio for low vs. high neighborhood education.

Conclusions: Researchers in eco-epidemiology should (i) develop a comprehensive, theory-based model of neighborhood determinants of participation in their study, (ii) investigate resulting biases for the environment–health associations, and (iii) check that unexplained geographic variations in participation do not bias these environment–health relationships.

Over the past 15 years, there has been a considerable development in the literature on neighborhood effects on health.[1-5] Cohort studies are typically used to investigate associations between neighborhood characteristics and health. However, such analyses suffer from a number of biases, including those related to selective participation in cohort studies,[6-8] which may distort the estimated associations between environmental exposures and health.[9]

As detailed in Online Appendix 1–A1, many selective participation biases may be formulated in terms of collider bias.[10-12] When the environmental exposure and the outcome, or factors affecting the exposure or the outcome, have causal effects on study participation, participation intervenes as a collider (i.e., a variable with at least two arrows pointing into it[11,13]). In these cases, conditioning on participation (in restricting the analysis to participants) can either generate an association between the environmental exposure and the outcome that does not exist in the source population or spuriously strengthen or weaken an existing association (see Online Appendix 1–A1).[10,14] Because differential participation rates and loss of follow-up are observed even in epidemiologic cohorts recruited through random sampling,[7,15] researchers investigating neighborhood effects should systematically investigate neighborhood determinants of study participation.[10]

Our first aim was to develop a comprehensive, theory-based model of neighborhood determinants of participation in the RECORD Cohort Study (see Online Appendix 1–A2). Our second aim was to examine whether neighborhood effects on study participation biased the associations between neighborhood socioeconomic variables and type 2 diabetes examined in this cohort (only few previous studies investigated relationships between neighborhood characteristics and diabetes[16]).

Importantly, biases in the environment–diabetes association of interest may result either from the influence of identified neighborhood characteristics on study participation or from the effects

of unidentified neighborhood factors on participation (as illustrated in Figure 1). We suggest that the neighborhood-level random effect of a model for study participation may be used to capture residual geographic variations in participation and control for its biasing effects. Building on Heckman selection models (see Online Appendix 1–E6), we attempt to correct some of the selective participation biases through the joint modeling of neighborhood determinants of participation and neighborhood determinants of diabetes.

METHODS

Population

Our investigation of the neighborhood determinants of study participation relied on two distinct databases: (i) the RECORD Study database for the number of RECORD participants per neighborhood and their sociodemographic characteristics, and (ii) the 1999 Census for the number of residents per neighborhood and their characteristics (denominators in the analyses). Our study of the neighborhood correlates of diabetes was based on the RECORD Cohort.

The RECORD Cohort: 7292 participants were recruited between March 2007 and February 2008. The participants were beneficiaries of the French National Health Insurance System for Salaried Workers, which offers a free medical examination every 5 years to all working and retired employees and their families (corresponding to 95% of the population of the Paris Ile-de-France region; see Online Appendix 1–D1). Participants were recruited without *a priori* sampling during these 2-hour-long preventive checkups conducted by the Centre d'Investigations Préventives et Cliniques in 4 of its health centers, located in the Paris Ile-de-France region (Paris, Argenteuil, Trappes, and Mantes-la-Jolie). Eligibility criteria were as follows: age 30 to 79 years; ability to fill out study questionnaires; and residence in one of the 10 (out of 20) administrative divisions of Paris or 122 municipalities of the metropolitan area selected *a priori* (corresponding

to a population of 5.2 million inhabitants in the 1999 Census). Importantly,[15] among people presenting at the health centers who were eligible based on age and residence, 10.9% were not selected for participation because of linguistic or cognitive difficulties in filling out study questionnaires. Of the persons selected for participation, 83.6% accepted to participate and completed the data collection protocol. Due to missing information, the available sample size was 7233 for study participation and 6876 for diabetes.

All participants underwent physical examination and filled out questionnaires. Participants were geocoded with accuracy based on their residential address in 2007–2008. Research assistants rectified all incorrect or incomplete addresses with the participants by telephone. Extensive investigations with local Departments of Urbanism were conducted to complete the geocoding. Spatial coordinates and geographic codes of street, block, and block group were searched for each participant. Precise coordinates and block group codes were identified for 100% of the participants. The study protocol was approved by the French Data Protection Authority.

The 1999 Population Census: The last available census, from 1999, was used for population denominators. A cross-tabulation provided the number of residents per age group, gender, and education level for each neighborhood.

Individual and neighborhood measures

Analyses of study participation

The following individual characteristics were divided into the same categories in the Population Census and in the RECORD database: age (30–39; 40–59; and 60 years or older), gender, and education level (no education; secondary school and lower tertiary education; and higher tertiary education).

Neighborhoods were defined as census block groups (IRIS areas in France). These were determined from the 1999 Census so as to be relatively homogeneous in sociodemographic and housing characteristics. Overall, 2218 neighborhoods were represented in the dataset matching the Population Census to the RECORD database. Fewer neighborhoods were represented in the RECORD database itself (1882 neighborhoods for the analyses on diabetes), because no participants were recruited in several of the neighborhoods in the study territory. The median number of residents in the 2218 neighborhoods was 2264 in 1999 (interquartile range: 1959, 2686). The median number of participants per neighborhood was 3 (interquartile range: 1, 5). Neighborhood median area size was 0.16 km² (interquartile range: 0.08, 0.35).

The following variables were considered at the neighborhood level: distance to the closest examination center; proportion of residents with a high education; median income; proportion of low income residents not paying taxes; proportion of the active population looking for work; proportion of residents receiving social benefits; mean dwelling value; population density; proportion of the area covered by buildings; mean building height; number of different public transportation lines accessible in the neighborhood; density of services; specialty care physician to primary care physician ratio; and an ecometric variable[17] for the degree of deterioration of the social/physical environment. Full details on these neighborhood variables and on hypotheses regarding their possible effects on study participation are reported in Online Appendix 1–C1 and 1–A3. All environmental variables were divided into quartiles.

Analyses of diabetes

Biological parameters were measured under fasting conditions. Diabetes was defined as fasting blood glucose ≥ 126 mg/dl or taking antidiabetic medication.

The following individual variables (described in Online Appendix 1–E2) were considered as possible correlates of diabetes: age and age^2, gender, marital status, education, and perceived financial strain. Three separate neighborhood variables (described in Online Appendix 1–C1) were used to characterize neighborhood socioeconomic position: the proportion of residents with a high education; median income; and mean dwelling value (see Online Appendix 1–B for hypotheses of neighborhood socioeconomic effects on diabetes).

Statistical methods

Models for study participation

In the analyses of study participation, the outcome was the number of RECORD participants (ranging from 0 to 16) in each individual sociodemographic stratum (based on age, gender, and education) of each neighborhood from the preselected municipalities. We specified a Poisson-distributed error and a log link function. The logarithm of neighborhood population in the corresponding sociodemographic stratum in the 1999 Census was specified as the offset. Geographic variations in the rate of study participation were taken into account by including a neighborhood random effect into the model.

To assess spatial autocorrelation in study participation, we estimated the Moran's I statistic for the neighborhood random effect of the model. In the absence of spatial autocorrelation, the Moran's I statistic has a small negative expectation when applied to regression residuals.[18] To investigate whether spatial correlation decreased with increasing distance between locations, we computed Moran's Is separately for neighborhoods less than 2,000 m apart, for those 2,000–3,999 m apart, those 4,000–5,999 m apart, and so forth.[19]

After estimating a model only adjusted for age and gender, we included individual education and the neighborhood variables into the model, only retaining those contextual variables that

were independently associated with participation. We explored cross-level interactions existing between individual-level education and neighborhood variables. As recently recommended,[20] after testing a model incorporating a product term of ordinal variables for individual education and the neighborhood variable, we estimated a model with a 12-category variable combining categories of individual education and of the neighborhood variable (allowing us to examine whether there was an interaction on either the additive or the multiplicative scale).

As reported in Online Appendix 1–F, we conducted a complementary analysis to distinguish between selection processes at different stages, i.e., separate contextual influences on the rate of people going for a health checkup and contextual influences on study participation among subjects who went for the checkup.

Models for diabetes

As detailed in Online Appendix 1–E2, we developed a multilevel logistic model for diabetes, testing a number of individual and neighborhood sociodemographic explanatory variables. To identify potential participation-related collider biases, first we examined whether some of the neighborhood determinants of study participation were associated with diabetes. We then extracted the median of the posterior distribution of the random effect for each neighborhood from the model on study participation, and used this random effect divided into quartiles as an explanatory variable to assess whether or not residual geographic variations in study participation were associated with diabetes.

However, the random effect capturing residual geographic variations in participation is not a directly observed quantity, but rather a model estimate implying uncertainty. To account for this uncertainty when estimating the association between residual geographic variations in study participation and diabetes, we used a Markov chain Monte Carlo approach to simultaneously

estimate the model for the neighborhood determinants of study participation and the model for diabetes. In this joint modeling, at each iteration of the chain, the current values of the neighborhood random effect for study participation (different from one iteration to the next) are inserted as an explanatory variable in the model for diabetes, permitting the associations between neighborhood socioeconomic variables and diabetes to be adjusted more accurately for the somewhat uncertain variable on the rate of participation.

All models were estimated with Markov chain Monte Carlo simulation using Winbugs 1.4.3.[21] All details on our estimation strategy are reported in Online Appendix 1–E3 to 1–E5 and the Winbugs code for all models is reported in Online Appendix 2.

RESULTS

Models for study participation

A multilevel model adjusted for age and gender revealed important between-neighborhood variations in study participation. Based on the between-neighborhood variance [variance = 0.21; 95% credible interval (CI): 0.18, 0.25], the rate of participation was 2.90 times higher (95% CI: 2.67, 3.15) for the 25% of all residents in neighborhoods with the highest rates of participation compared with the 25% of all residents in neighborhoods with the lowest rates.[3-4,22] As shown with the Moran's I (Figure 2), spatial autocorrelation in study participation was observed over a large range but was modest in magnitude. The correlation decreased with increasing distance between neighborhoods and vanished for neighborhoods 12 km or further apart.

The distribution of the study participants and total population according to the individual and neighborhood characteristics is reported in Online Appendix 1–D2. A model containing individual and neighborhood variables indicated a markedly higher rate of study participation for

individuals with a high education attainment (Table 1). The rate of participation was lower for people residing far from the study center. Regarding socio-environmental variables, the rate of study participation was higher in both high median income and high mean dwelling value neighborhoods after controlling for individual education. By contrast, the rate of participation was higher in neighborhoods with a high proportion of the active population looking for work (see discussion for an interpretation related to the recruitment strategy of the health centers).

Regarding physical environmental variables, independent associations indicated higher rates of study participation in neighborhoods with a low proportion of the area covered by buildings and a low mean building height. Contrary to our expectations, the ecometric variable representing the deterioration of the social/physical environment was not associated with participation.

Pearson correlations between these neighborhood variables were moderate with a few exceptions (Online Appendix 1–C2).

Product terms between individual education and neighborhood variables coded as ordinal variables indicated that there was an interaction between the effects of individual education and distance to the center on the multiplicative scale. However, the model reported in Table 2 revealed that the negative effect of distance on study participation was stronger among individuals with low education levels when assessed on the multiplicative scale; whereas the effect of distance was larger in the high education group when the interaction was assessed on the additive scale.

As detailed in Online Appendix 1–F, complementary analyses conducted with individuals nested within municipalities confirmed that distance to the center and area indicators of socioeconomic position and density were associated with going for the health checkup but were not associated (or only very weakly) with study participation among persons who were at the examination center for the health checkup.

In the final model for study participation, the between-neighborhood variance was reduced to 0.12 (95% CI: 0.09, 0.14). As shown in Figure 2, spatial autocorrelation in study participation was to a large extent explained by the individual and neighborhood variables introduced into the model.

Models for diabetes

As shown in Table 3 (first column), a low neighborhood education was associated with slightly higher odds of diabetes, after controlling for individual education and self-reported financial strain (see Online Appendix 1–E2 for details on the construction of the model). Apart from neighborhood education, none of the neighborhood determinants of study participation (distance to the center, income, dwelling value, proportion looking for work, building density and height) showed associations with diabetes. Therefore, there was no need to adjust the model on diabetes for these neighborhood factors to remove participation-related collider biases.

The neighborhood-level random effect of the final model for study participation, capturing residual geographic variations in participation, was associated with the odds of diabetes, which were slightly higher in high-participation areas (Table 3, column 2). The neighborhood random effect of the final model for participation showed almost no correlation with neighborhood education in the general population (r = –0.004; 95% confidence interval: –0.005, –0.002; n = 3.1 million). However, as expected from Figure 1, this random effect was negatively associated with neighborhood education in the sample of participants (r = –0.14; 95% confidence interval: –0.17, –0.12; n = 7233). Compared to the general population, the relationship between the study participation-random effect and neighborhood education was pulled into the negative in the sample of participants because, if participation in the study is not caused by residing in a socially advantaged neighborhood, then it is likely that another cause of participation is present, e.g.,

residing in one of these unspecified high-participation areas (identified from the participation random effect).

Due to this correlation, it is probably relevant to take into account residual geographic variations in study participation when estimating the association between neighborhood education and diabetes. As expected from Figure 1, the association between neighborhood education and diabetes was slightly reduced when the median of the posterior distribution of each neighborhood's participation-random effect was introduced as a predictor in the model for diabetes (the change in effect size between columns 1 and 2 of Table 3 was extremely minimal but was in the expected direction).

However, as noted above, the uncertainty associated with the random effect of the participation model would need to be taken into account in our adjustment of the model for diabetes. To do so, we relied on the Markov chain Monte Carlo framework to estimate the model for the neighborhood determinants of study participation jointly with the model for diabetes (inserting the random effect of the first model as an explanatory variable in the second one). As shown in Table 4, in this joint model for participation and diabetes, the neighborhood random effect of the model for study participation was associated with the odds of diabetes. The log prevalence odds ratio for diabetes in low vs. high education neighborhoods was 18% lower in the joint model (prevalence odds ratio = 1.44, 95% CI: 0.98, 2.13) than in the model of Table 3 (prevalence odds ratio = 1.56, 95% CI: 1.06, 2.31) that does not control for residual geographic variations in study participation.

DISCUSSION

We found that a number of neighborhood factors related to the socioeconomic and physical environments were associated with participation in the RECORD Cohort Study, suggesting that

participation biases may not only depend on individual characteristics but also on neighborhood features. Investigating associations between neighborhood socioeconomic variables and diabetes, we found that residual geographic variations in the rate of study participation were associated with diabetes. We attempted to correct the resulting bias in the relatively weak neighborhood education–diabetes association that was observed through the joint modeling of the determinants of study participation and diabetes.

Strengths and limitations

Strengths of the present study include the original research design which allowed us to investigate individual/neighborhood determinants of participation in a cohort study, the large number of environmental correlates of participation that we considered, the fact that residual random geographic variations in participation were conceptualized as a potential source of participation-related collider bias, and the joint modeling framework implemented for bias correction.

One limitation of the participation analysis is the mismatch between the Census and RECORD Study data. Discrepancies between numerators and denominators include (i) the mismatch between the Census date (1999) and the RECORD recruitment dates (2007–2008), (ii) and the fact that individuals eligible for the health checkup had to be affiliated with the French National Health Insurance System for Salaried Workers, which corresponds to 95% of the total Census population. It is unlikely, however, that this mismatch could have sufficiently affected denominators of the participation rate so as to produce the observed associations with study participation. Another critical limitation is our inability to examine whether blood glucose or diabetes influenced study participation (we had information on diabetes neither for the general

population nor for the persons who came to the health centers but did not participate in the study).

Main findings

Neighborhood influences on study participation

There were 3 different selection stages in our recruitment strategy. First, populations attending the health centers are specific. Second, particular participants were excluded by the staff (because of linguistic or cognitive limitations). Third, those who accepted to participate were also possibly specific. Analyses reported in the main article amalgamated the 3 sources of selection, while those reported in Online Appendix (1–F) distinguished between (i) the first selection stage and (ii) the second and third stages amalgamated together.

We found that the longer the road network distance was to the closest health center, the lower the rate of study participation. As expected, complementary analyses confirmed that distance to the center predicted attendance for the health checkup but not study participation among people who were at the center for the health checkup (see Online Appendix 1–F). Notably, the inhibiting effect of distance was more acute among persons with low education levels when assessed on the multiplicative scale, but weaker among these low educated participants when the interaction was assessed on the additive scale (due to the fact that the basal rate of participation was much higher in educated than in non-educated participants). Due to the absence of strong theoretical guidance to decide which of the additive or multiplicative definition of effects should be considered to gauge the interaction of interest in this particular case, it seems difficult to firmly conclude that the distance effect on participation was stronger or weaker in low than in high educated populations.

Independent of individual education effects, two mutually adjusted neighborhood effects (resulting from median income and mean dwelling value) indicated a lower rate of participation for residents of deprived neighborhoods.[15] Complementary analyses (see Online Appendix 1–F) showed that low individual education did not strongly decrease the rate of people going for a health checkup, but was very strongly associated with low study participation among people who had come to the health centers for a checkup (perhaps reflecting a low interest in scientific studies[6,23] and exclusion from the study because of linguistic or cognitive difficulties in filling out questionnaires among low socioeconomic groups). By contrast, a low neighborhood socioeconomic status was only associated with slightly lower rates of participation among people seen at the health centers, but was associated with a markedly lower rate of attendance at the centers (perhaps reflecting the spatial isolation of deprived neighborhoods, their lack of efficient public transportation, their residents' tendencies to rely on local resources, and collective norms that do not promote preventive healthcare). Importantly, it is possible that no neighborhood effect was detected on the acceptance to participate among persons seen at the health centers because these people were a selected set of individuals who made significant efforts to attend the health centers. We cannot exclude that neighborhood effects would have been noted if a less selected population, contacted in its residential environment, were invited to participate.

By contrast, after adjustment, a high proportion of residents looking for work was associated with a *higher* rate of participation. As this variable reflects socioeconomic instability, a possible explanation is related to the specific recruitment targets of the participating health centers. Indeed, 3 of the 4 recruiting centers were set up in highly deprived areas specifically to reach patients with unstable economic resources.

Neighborhood built surface and average building height consistently indicated that higher building densities were associated with lower rates of study participation.[15] In the absence of

more convincing hypotheses, we can only speculate that residents of sparsely populated neighborhoods may have specific health-related attitudes encouraging them to attend preventive health examination centers.

Selective study participation as a source of bias in the neighborhood–diabetes association

Online Appendix 1–A describes a number of situations in which environmental influences on study participation could bias environment–health associations. In our case, if building density (a determinant of participation) was a cause of diabetes, we would have to adjust our association between neighborhood education and diabetes for density. This is because, even in the absence of a relationship between neighborhood education and building density in the general population, conditioning on participation would generate an association between them.

In our analyses, none of the identified neighborhood determinants of study participation could bias the relatively weak association identified between neighborhood education and diabetes. One of the original ideas of the study was to rely on the neighborhood random effect of the participation model to (i) capture the effects of unidentified neighborhood characteristics on study participation, (ii) examine whether these residual geographic variations in participation were associated with diabetes, and (iii) adjust the health model to remove a possible selective participation bias. By definition, this approach is not hypothesis-driven and does not need to be (we have no idea of the nature of neighborhood influences on participation captured by the random effect and why the latter was associated with diabetes). Overall, even if the bias correction only lead to a relatively weak change in the estimate of interest, our example illustrates that this strategy may enable to correct participation-related collider biases that are not easily identifiable.

Implications for future investigations

Our study shows that it is feasible to investigate neighborhood determinants of participation in cohort studies. Of course, neighborhood-related selection may be much weaker when recruitment is based on *a priori* randomization and invitation of selected participants, and still weaker when participants are further surveyed and examined at home or nearby. However, relying on a randomized sample is not sufficient (due to selective non-participation and attrition), and eco-epidemiologists, in addition to minimizing selection effects, should develop a comprehensive knowledge of the neighborhood determinants of participation in their study.

Overall, the general recommendations we make for ourselves, recommendations which may also be relevant for others, are as follows: (i) we will extend our analyses of the neighborhood determinants of participation in the RECORD Study; (ii) we will rely on this comprehensive list of neighborhood determinants of participation to test their association with health outcomes in a search of participation-related collider biases; and (iii) we will rely on the proposed joint modeling framework to verify that unexplained geographic variations in study participation do not bias the environment–health associations of interest.

References

1. Riva M, Gauvin L, Barnett TA. Toward the next generation of research into small area effects on health: a synthesis of multilevel investigations. *J Epidemiol Community Health*. 2007;61:853-861.
2. Chaix B. Geographic Life Environments and Coronary Heart Disease: A Literature Review, Theoretical Contributions, Methodological Updates, and a Research Agenda. *Annu Rev Public Health*. 2009;30:81-105.
3. Chaix B, Rosvall M, Merlo J. Recent increase of neighborhood socioeconomic effects on ischemic heart disease mortality: a multilevel survival analysis of two large Swedish cohorts. *Am J Epidemiol*. 2007;165:22-26.
4. Chaix B, Rosvall M, Merlo J. Neighborhood socioeconomic deprivation and residential instability: effects on incidence of ischemic heart disease and survival after myocardial infarction. *Epidemiology*. 2007;18:104-111.
5. Chaix B, Rosvall M, Merlo J. Assessment of the magnitude of geographic variations and socioeconomic contextual effects on ischaemic heart disease mortality: a multilevel survival analysis of a large Swedish cohort. *J Epidemiol Community Health*. 2007;61:349-355.
6. Goldberg M, Chastang JF, Leclerc A, et al. Socioeconomic, demographic, occupational, and health factors associated with participation in a long-term epidemiologic survey: a prospective study of the French GAZEL cohort and its target population. *Am J Epidemiol*. 2001;154:373-84.
7. Nohr EA, Frydenberg M, Henriksen TB, Olsen J. Does low participation in cohort studies induce bias? *Epidemiology*. 2006;17:413-8.

8. Lissner L, Skoog I, Andersson K, et al. Participation bias in longitudinal studies: experience from the Population Study of Women in Gothenburg, Sweden. *Scand J Prim Health Care*. 2003;21:242-7.

9. Greenland S. Response and follow-up bias in cohort studies. *Am J Epidemiol*. 1977;106:184-7.

10. Hernan MA, Hernandez-Diaz S, Robins JM. A structural approach to selection bias. *Epidemiology*. 2004;15:615-25.

11. Fleischer NL, Diez Roux AV. Using directed acyclic graphs to guide analyses of neighbourhood health effects: an introduction. *J Epidemiol Community Health*. 2008;62:842-6.

12. Cole SR, Platt RW, Schisterman EF, et al. Illustrating bias due to conditioning on a collider. *Int J Epidemiol*. 2010;39:417-20.

13. Chaix B, Leal C, Evans D. Neighborhood-level Confounding in Epidemiologic Studies: Unavoidable Challenges, Uncertain Solutions. *Epidemiology*. 2010;21:124-127.

14. Greenland S. Quantifying biases in causal models: classical confounding vs collider-stratification bias. *Epidemiology*. 2003;14:300-6.

15. Oakes JM, Forsyth A, Hearst MO, Schmitz KH. Recruiting Participants for Neighborhood Effects Research: Strategies and Outcomes of the Twin Cities Walking Study. *Environ Behav*. 2009;41:787-805.

16. Leal C, Chaix B. The influence of geographic life environments on cardiometabolic risk factors: a systematic review, a methodological assessment and a research agenda. *Obes Rev*. 2010; in press.

17. Chaix B, Lindstrom M, Merlo J, Rosvall M. Neighbourhood social interactions and risk of acute myocardial infarction. *J Epidemiol Community Health*. 2008;62:62-8.

18. Odland J. *Spatial autocorrelation*. Newbury Park, CA: Sage Publications; 1988.

19. Chaix B, Merlo J, Subramanian SV, Lynch J, Chauvin P. Comparison of a spatial perspective with the multilevel analytic approach in neighborhood studies: the case of mental and behavioral disorders due to psychoactive substance use in Malmö, Sweden, 2001. *Am J Epidemiol*. 2005;162:171-182.

20. Kaufman JS. Interaction reaction. *Epidemiology*. 2009;20:159-60.

21. Smith AFM, Roberts GO. Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *J R Stat Soc Ser B Stat Methodol*. 1993;55:3-23.

22. Merlo J, Chaix B, Ohlsson H, et al. A brief conceptual tutorial of multilevel analysis in social epidemiology – using measures of clustering in multilevel logistic regression to investigate contextual phenomena. *J Epidemiol Community Health*. 2006;60:290-297.

23. Turrell G, Patterson C, Oldenburg B, Gould T, Roy MA. The socio-economic patterning of survey participation and non-response error in a multilevel study of food purchasing behaviour: area- and individual-level characteristics. *Public Health Nutr*. 2003;6:181-9.

Figure legends

Figure 1

Unidentified neighborhood characteristics influencing participation in the study, if also associated with the outcome (type 2 diabetes), may bias the association of interest between neighborhood average education and diabetes. The dashed line represents the association generated by restricting the analyses to participants. Following Hernán (Epidemiology 2004;15:615-25), the rectangle around participation indicates that the analyses condition on participation. The plus and minus signs indicate the direction of the associations observed in the data.

Figure 2

Moran's I statistics for neighborhood-level residuals of multilevel models for participation in the RECORD Cohort Study, computed separately for pairs of neighborhoods less than 2,000 m apart, 2,000–3,999 m apart, 4,000–5,999 m apart, etc. The initial model only included age and gender; individual education and neighborhood factors were introduced in the final model. Bars represent 95% credible intervals.

TABLE 1. Rate ratios (RR) for the associations between individual/neighbourhood characteristics and participation in the RECORD Cohort Study, as estimated from a multilevel Poisson model (all effects adjusted for each other)

	RR (95% CI)
Age (vs. 30–39 years)	
40–59 years	1.84 (1.74, 1.96)
60 years and over	1.37 (1.27, 1.47)
Men (vs. women)	2.00 (1.90, 2.10)
Individual education level (vs. low)	
Medium	1.90 (1.74, 2.08)
High	4.25 (3.87, 4.67)
Distance to the center (vs. high)	
Mid-high	1.19 (1.09, 1.30)
Mid-low	1.45 (1.32, 1.58)
Low	1.75 (1.60, 1.91)
Median income (vs. low)	
Mid-low	1.20 (1.09, 1.32)
Mid-high	1.29 (1.14, 1.45)
High	1.39 (1.20, 1.60)
Mean dwelling value (vs. low)	
Mid-low	1.10 (1.00, 1.21)
Mid-high	1.11 (1.00, 1.24)
High	1.23 (1.09, 1.39)
Proportion of the active population looking for work (vs. low)	
Mid-low	1.01 (0.93, 1.10)
Mid-high	1.18 (1.06, 1.31)
High	1.31 (1.15, 1.47)
Proportion of the area covered by buildings (vs. high)	
Mid-high	1.13 (1.03, 1.23)
Mid-low	1.26 (1.14, 1.39)
Low	1.37 (1.23, 1.51)
Mean building height (vs. high)	
Mid-high	1.11 (1.03, 1.21)
Mid-low	1.27 (1.16, 1.39)
Low	1.27 (1.15, 1.40)

TABLE 2. Rate ratios (RR) for the association between combined categories of individual education and distance to the closest center on the one hand, and participation in the RECORD Cohort Study on the other hand, adjusted for age, gender and neighborhood variables, as estimated from a multilevel Poisson model[a]

	RR	**95% CI**
Low individual education		
High distance to the center	Ref.	
Mid-high distance to the center	1.23	(0.94, 1.61)
Mid-low distance to the center	1.56	(1.21, 2.03)
Low distance to the center	2.75	(2.19, 3.47)
Intermediate individual education		
High distance to the center	2.32	(1.93, 2.83)
Mid-high distance to the center	2.60	(2.15, 3.19)
Mid-low distance to the center	3.27	(2.71, 4.02)
Low distance to the center	4.06	(3.35, 4.97)
High individual education		
High distance to the center	5.28	(4.31, 6.54)
Mid-high distance to the center	6.61	(5.44, 8.18)
Mid-low distance to the center	7.49	(6.15, 9.24)
Low distance to the center	8.04	(6.59, 9.91)

[a]On the multiplicative scale, the rate ratio for participation between people living nearby and far from the closest health center was 2.75 (2.75/1) in the low education group, 1.75 (4.06/2.32) in the intermediate education group, and 1.52 (8.04/5.28) in the high education group. In contrast, on the additive scale, for a basal rate of participation equal to R, the effect of distance was 1.75R in the low education group, 1.74R in the intermediate education group, and 2.76R in the high education group.

TABLE 3. Associations between individual and neighborhood characteristics and the odds of diabetes, as estimated from multilevel logistic models (all effects adjusted for each other), before and after controlling for residual geographic variations in the rate of study participation, RECORD Cohort Study, n = 6876

	Before adjustment	**After adjustment**
	POR[a] (95% CI)	**POR (95% CI)**
Age (1-year increase)	1.24 (1.07, 1.38)	1.25 (1.13, 1.41)
Age square	1.00 (1.00, 1.00)	1.00 (1.00, 1.00)
Male vs. female	1.38 (1.05, 1.84)	1.39 (1.06, 1.86)
Living alone vs. cohabiting	0.97 (0.72, 1.30)	0.99 (0.73, 1.32)
Individual education (vs. high)		
Medium	1.40 (1.04, 1.89)	1.39 (1.02, 1.87)
Low	1.94 (1.26, 2.92)	1.91 (1.24, 2.88)
Perceived financial strain (vs. not)	1.52 (1.07, 2.14)	1.53 (1.07, 2.16)
Neighborhood education (vs. high)		
Mid-high	1.05 (0.70, 1.56)	1.02 (0.68, 1.52)
Mid-low	1.19 (0.80, 1.75)	1.17 (0.79, 1.73)
Low	1.56 (1.06, 2.31)	1.50 (1.01, 2.23)
Neighborhood random effect for study participation (vs. low)		
Mid-low		1.19 (0.81, 1.77)
Mid-high		1.31 (0.89, 1.93)
High		1.58 (1.09, 2.33)

[a]POR, prevalence odds ratio.

TABLE 4. Joint modeling (i) of the associations between individual and neighbourhood characteristics and participation in the RECORD Study, and (ii) of the associations between individual and neighbourhood characteristics and the odds of diabetes (all effects in each model adjusted for each other)

	Exp(β)[a]	95% CI
Model for participation in the RECORD Study		
Age (vs. 30–39 years)		
40–59 years	1.84	(1.73, 1.95)
60 years and over	1.36	(1.27, 1.47)
Men (vs. women)	2.00	(1.90, 2.10)
Individual education × distance to the center		
Low individual education		
High distance	(Ref.)	
Mid-high distance	1.25	(0.96, 1.63)
Mid-low distance	1.58	(1.22, 2.04)
Low distance	2.77	(2.21, 3.48)
Intermediate individual education		
High distance	2.36	(1.95, 2.86)
Mid-high distance	2.63	(2.17, 3.22)
Mid-low distance	3.30	(2.73, 4.03)
Low distance	4.09	(3.38, 5.00)
High individual education		
High distance	5.37	(4.38, 6.61)
Mid-high distance	6.71	(5.49, 8.25)
Mid-low distance	7.55	(6.20, 9.27)
Low distance	8.10	(6.63, 9.94)
Median income (vs. low)		
Mid-low	1.19	(1.08, 1.32)
Mid-high	1.28	(1.13, 1.45)
High	1.39	(1.20, 1.61)
Mean dwelling value (vs. low)		
Mid-low	1.10	(1.00, 1.21)
Mid-high	1.12	(1.00, 1.24)
High	1.24	(1.10, 1.40)
Proportion of the active population looking for work (vs. low)		
Mid-low	1.01	(0.93, 1.10)
Mid-high	1.18	(1.07, 1.32)
High	1.31	(1.16, 1.49)
Proportion of the area covered by buildings (vs. high)		
Mid-high	1.12	(1.03, 1.21)
Mid-low	1.24	(1.13, 1.37)
Low	1.34	(1.21, 1.49)
Mean building height (vs. high)		
Mid-high	1.11	(1.02, 1.20)
Mid-low	1.25	(1.15, 1.37)
Low	1.26	(1.15, 1.38)

Model for diabetes		
Age (1-year increase)	1.25	(1.12, 1.37)
Age square	1.00	(1.00, 1.00)
Male vs. female	1.39	(1.05, 1.85)
Living alone vs. cohabiting	0.98	(0.73, 1.31)
Individual education (vs. high)		
Medium	1.39	(1.03, 1.88)
Low	1.88	(1.23, 2.84)
Perceived financial strain (vs. not)	1.52	(1.07, 2.12)
Neighborhood education (vs. high)		
Mid-high	1.01	(0.68, 1.48)
Mid-low	1.15	(0.78, 1.69)
Low	1.44	(0.98, 2.13)
Neighborhood random effect for study participation (continuous)	2.90	(1.39, 6.39)

[a]Parameters reported for the model on study participation are rate ratios; parameters reported for the model on diabetes are prevalence odds ratios.

Figure 1

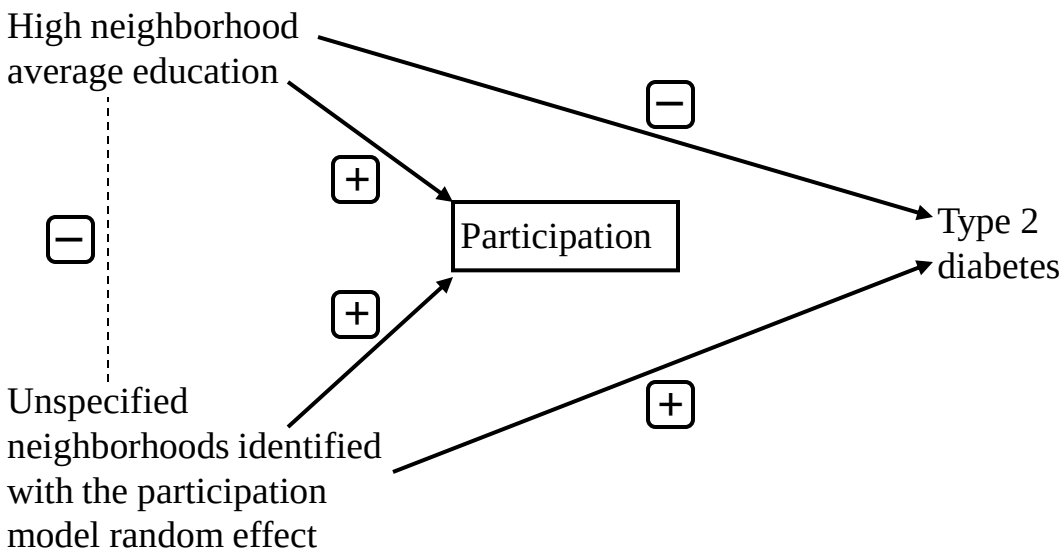
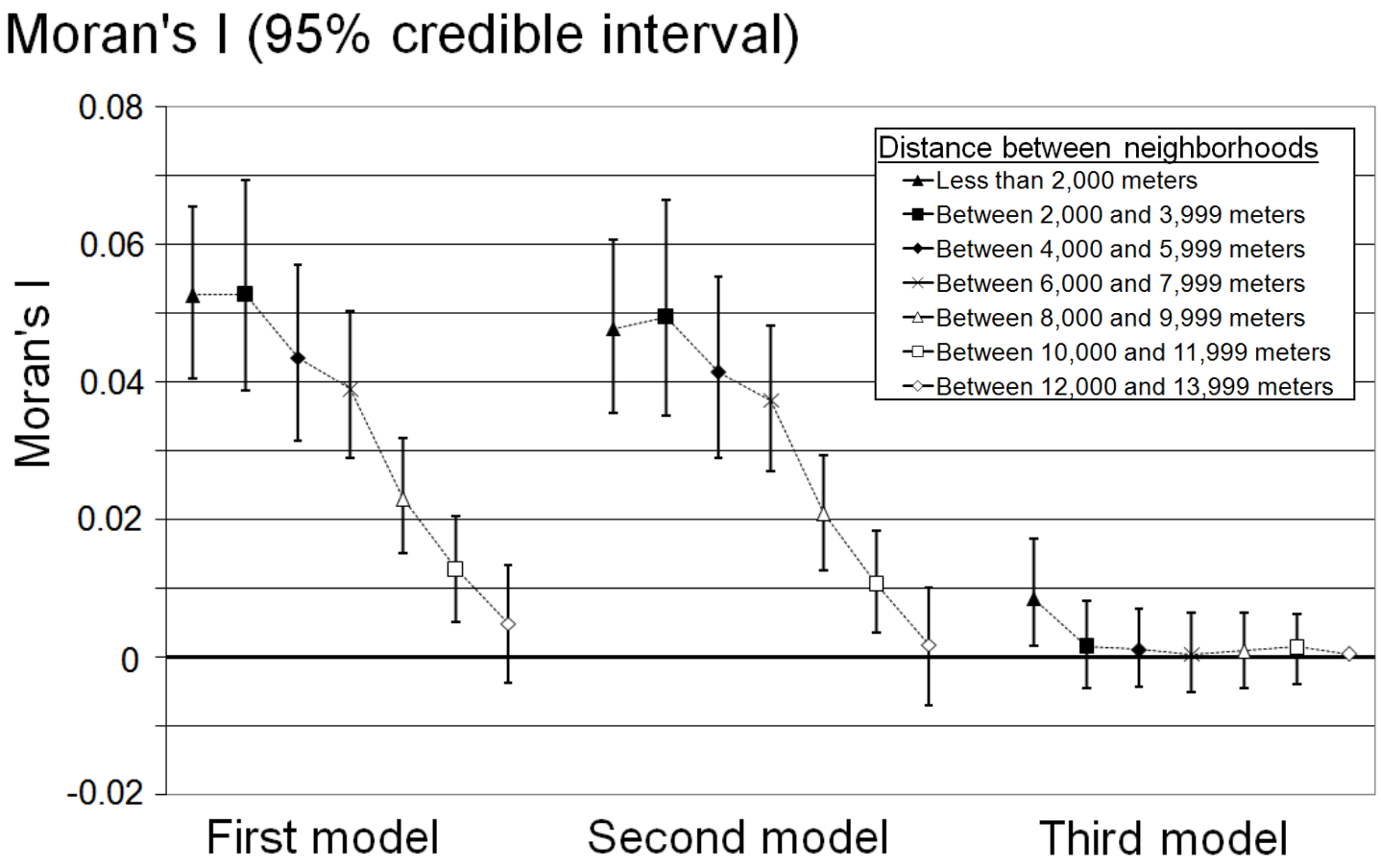

High neighborhood average education
Participation
Type 2 diabetes
Unspecified neighborhoods identified with the participation model random effect
−
+
−
+
+

Figure 2


Moran's I (95% credible interval)
Moran's I
0.08
0.06
0.04
0.02
0
-0.02
First model
Second model
Third model
Distance between neighborhoods
Less than 2,000 meters
Between 2,000 and 3,999 meters
Between 4,000 and 5,999 meters
Between 6,000 and 7,999 meters
Between 8,000 and 9,999 meters
Between 10,000 and 11,999 meters
Between 12,000 and 13,999 meters

A joint modeling of neighborhood effects on participation in the RECORD Cohort Study and neighborhood effects on type 2 diabetes: bias assessment and correction

Online Appendix 1

A – Rationale: Neighborhood effects on study participation and related biases

A1 – Biases resulting from neighborhood effects on study participation

Recent articles have emphasized that many selective participation biases may be formulated in terms of collider biases.[1-2] In Figures S1-A and S1-B of the present appendix, we develop hypothetical scenarios of neighborhood effects on obesity that only have an illustrative purpose and are not related to the empirical analyses developed in the main text. Figure S1-A is interested in the association between fast-food outlet density and obesity. Study participation intervenes as a collider (i.e., a variable with two arrows pointing into it[2]) because both low neighborhood socioeconomic position and individual obesity decrease the odds of study participation. By conditioning on participation (by restricting the analysis to participants), we generate an association between neighborhood socioeconomic position and obesity that does not exist in the population: if a participant does not reside in a high socioeconomic position neighborhood (a cause of participation), then it is more likely that another cause of participation is present, i.e., better health status.[1] In Figure S1-A, such a collider bias may affect the estimated association between fast-food outlet density and obesity.

Different scenarios involving neighborhood variables may result in selective participation collider biases. These scenarios include cases in which both the main environmental exposure and the outcome are direct determinants of participation, cases as in Figure S1-A and S1-B in which only the exposure or the outcome is a determinant of participation (but in which a parent of the other variable is a determinant of participation), and cases in which neither the exposure nor the outcome influence participation (but parents of both variables do) as in the so-called M-bias.[3-4] We may expect selective participation biases to be the most threatening when both the exposure and the outcome are direct determinants of study participation, and the least threatening when only parents of these variables have a causal effect on participation. Overall, it is important to note that participation-based collider biases are possible even in the absence of any direct or indirect influence of health on participation (as in Figure S1-B of this appendix).

Fortunately, in Figures S1-A and S1-B, it is possible to close the backdoor path created by restriction to participants by estimating the effect of interest conditional on neighborhood socioeconomic position.[1] However, in situations where both the main exposure and the outcome are direct determinants of participation, there may be no alternative but to simply evaluate the direction and magnitude of bias.

Other scenarios for selective participation bias include cases where an environmental determinant of participation is a modifier of the association between the main environmental exposure and the outcome,[2] or cases where the main environmental exposure modifies the effect of health on participation.[5-6]

A2 – Why is it relevant to derive a comprehensive model of the neighborhood determinants of participation in a specific cohort study?

Our aim was to build a *comprehensive* model of the neighborhood determinants of participation in the RECORD Cohort Study, i.e., to identify the neighborhood determinants of participation as exhaustively as possible, based on hypotheses of factors that may influence participation. As for the modeling of neighborhood determinants of health outcomes, it is important to follow a hypothesis-driven approach to investigate neighborhood determinants of participation in a cohort study (see section A3 below).

When investigating associations between neighborhood factors and health, it is perhaps less easy to rely on intuition and basic reasoning to identify selection biases resulting from neighborhood influences on study participation than it is to identify confounding biases

resulting from the effect of a neighborhood factor on the exposure and the outcome.[7] Identification of confounders requires knowledge of the causal relationships between the variables of interest (environmental exposure, outcome, and determinants of environmental exposure and outcome) whereas selection bias identification implies more abstract and methodological reasoning related to the process of construction of the study sample. Eco-epidemiologists are not used to paying attention to the neighborhood determinants of participation, which precludes the identification of participation-related collider biases.

In Figure S1-A of the present appendix, neighborhood socioeconomic position has no effect on the health outcome, but it nonetheless contributes to a collider bias because of its causal effect on the exposure of interest. Therefore, the range of environmental factors that could contribute, if associated with study participation, to biasing the association of interest is *a priori* very broad.

Accordingly, we argue that eco-epidemiologists should *a priori* obtain comprehensive knowledge of the neighborhood determinants of participation in their cohort study, as a guide in the identification of potential neighborhood-related participation collider biases that are all but intuitive. An exhaustive list of neighborhood determinants of participation would be useful to identify problematic cases where (i) the environmental exposure of interest, (ii) an environmental determinant of the exposure of interest, or (iii) an environmental determinant of the health outcome influences study participation.

A3 – Hypotheses of neighborhood effects on participation in the RECORD Cohort Study

It is important to rely on a relevant theoretical model of the determinants of study participation to investigate neighborhood influences on participation.

We assumed that the following contextual circumstances may be associated with a lower rate of participation in the RECORD Cohort Study: a longer distance to the center, a low socioeconomic profile (with a particular interest for a low education and low income that may reflect, respectively, a certain disinterest for preventive care and less consumerist attitudes), and a low density of services (which may also be associated with less pronounced consumerist attitudes). It was also hypothesized that a high degree of deterioration of the social/physical environment implies chronic difficulties that do not encourage preventive healthcare behavior.

In opposition to the main neighborhood socioeconomic effects generating higher participation rates in advantaged neighborhoods, we expected that markers of socioeconomic instability or high poverty might be associated with higher participation rates, because of the specific recruitment strategies of the healthcare centers involved in the study (which particularly try to offer preventive health examinations to subpopulations in situations of high poverty / socioeconomic instability).

We also considered the ratio of specialty to primary care physicians in the neighborhood. We assume that reciprocal causal influences may exist between this variable and an attitudinal variable reflecting the interest of populations for high quality care and preventive care. The latter variable may be associated with the rate of people attending health centers, thus participating in the RECORD Study, and the specialty to primary care physicians ratio may serve as a proxy to capture this effect.

B – Rationale: Hypotheses of neighborhood socioeconomic effects on type 2 diabetes

In the present section, we heavily rely on the published literature to enumerate various mechanisms through which the neighborhood socioeconomic environment may influence the development of type 2 diabetes.[8-11] Our hypothesis is that neighborhood socioeconomic status is a fundamental cause of disease that contributes to shape a number of more proximate environmental resources or exposures that have a direct effect on the development of type 2 diabetes.

A number of the hypothesized mechanisms are based on the fact that an increase in body mass index or waist circumference is important in the pathogenesis of diabetes. As authors have stated,[9] body mass index is rather proximal to insulin resistance in the causal chain leading from area features, through behavior, to insulin resistance.

It is commonly emphasized that residential environments may affect both diet and physical activity, which are two important risk factors for metabolic abnormalities.[9] Regarding diet, on the one hand, it has been shown that the neighborhood socioeconomic context strongly influences the degree of availability of healthy foods.[12] Authors have argued that food stores in poor neighborhoods are less likely to sell healthier items such as low-fat and high-fiber products.[10] On the other hand, it is hypothesized that the local availability of high-quality fruits and vegetables and of low-fat foods is an important determinant of a healthy diet.[9]

Regarding physical activity, walking destinations and opportunities for physical activity, such as parks and sport or recreational facilities, are important environmental resources to promote active living.[9,13] Previous literature has shown that, in a number of settings, there may be a lower availability of recreational and sport facilities in low-income neighborhoods.[14]

Overall, as authors have emphasized, diet and physical activity may be two of the proximate mechanisms through with neighborhood socioeconomic characteristics influence the development of the insulin resistance syndrome and incidence of type 2 diabetes.[8]

Moreover, authors have suggested that chronic stress may be related to the development of the insulin resistance syndrome through endocrine pathways.[8] As these authors indicate, sources of chronic stress (such as noise, violence, and poverty) are likely to vary across neighborhoods, and could be involved in linking residential environments to the development of the insulin resistance syndrome.[8,11]

Finally, some authors have hypothesized that environmental conditions such as dioxin, lead, or other toxic exposures may play a role in the association of adverse neighborhood conditions with the development of diabetes.[11]

In all the aforementioned hypotheses, neighborhood socioeconomic status is conceptualized as a fundamental cause that influences the risk of diabetes through other mediating environmental exposures or resources. However, it should be kept in mind that more direct effects of particular facets of the neighborhood socioeconomic environment are also possible, such as the direct effect of neighborhood average education and related capital of knowledge on diet and physical activity.

C – Definition of neighborhood variables and correlation between these variables

C1 – Definition of neighborhood variables

We considered a number of different neighborhood variables as possible determinants of participation in the RECORD Cohort Study.

Using ArcGIS 9.2 Network Analyst with street network data from IGN (National Geographic Institute), we determined the street network distance from each neighborhood centroid to the closest of the 4 study centers.

The following socioeconomic variables were defined at the neighborhood level: (i) the proportion of residents aged 15 or over with an upper tertiary education (1999 Census); (ii) area population density (the number of residents from the 1999 Census per km²); (iii) median income in 2005 and (iv) the proportion of low income individuals not paying taxes in 2005 (Tax Registry of DGI, General Directorate of Taxation); (v) the proportion of the active population looking for work in 2006 and (vi) the proportion of very low income, unemployed persons receiving social benefits in 2006 (ANPE, the National Employment Agency); and (vii) mean value of dwellings sold in 2003–2007 (Paris-Notaires).

Using ArcGIS, we defined the following neighborhood variables related to the physical and service environments: (i) the proportion of the area covered by buildings and (ii) mean building height in 2008 (from IGN); (iii) the number of different transportation lines (buses, trains, tramways) accessible in 2008 (STIF, the Authority for Public Transport in the Paris Region); (iv) the density of services per km² in 2005 including public and administration services, all types of public/private shops, entertainment facilities, etc. (Permanent Database of Facilities from INSEE, the National Institute of Statistics and Economic Studies); and (v) the ratio of specialty care to primary care physicians (determined from the Visiaurif-Santé of IAU-IdF).

Finally, a variable for deterioration of the social/physical environment was defined at the TRIRIS area level (i.e. areas merging approximately three IRIS neighborhoods). Following the ecometric approach,[15-17] we estimated a three-level (survey questions, individuals, TRIRIS areas) multilevel logistic model with the RECORD participants' answers to 6 questions about their neighborhood as the outcome. These questions were related to block face deterioration, insufficient maintenance of neighborhood facilities, presence of garbage and graffiti, incivilities, vandalism, and excessive noise from the neighbors. There were a high intra-individual correlation and intra-TRIRIS correlation in the answers to these questions, suggesting that the scale was psychometrically and ecometrically sound.[18] Multilevel models allowed us to aggregate at the individual level the information provided by each respondent and to combine the answers of the different individuals of the same neighborhood to construct indicators at the neighborhood level. As recommended, the multilevel model TRIRIS-level random effect was then used as an explanatory variable quantifying the degree of deterioration of the social/physical environment. In order to derive reliable environmental variables with this approach, it is necessary to have a sufficient number of individuals per neighborhood assessing their environment. That is why our ecometric measurement protocol was *a priori* conceived to be implemented, not at the most local neighborhood level, but at the level of slightly larger neighborhoods (TRIRIS areas).

As noted in the main article, our aim was to test whether 3 specific neighborhood socioeconomic variables were associated with type 2 diabetes: the proportion of neighborhood residents with an upper tertiary education, neighborhood median income, and neighborhood mean dwelling value (see description above).

C2 – Correlation between the neighborhood variables

When including multiple neighborhood variables in a regression model, it is important to check that these variables are not too correlated to separate their effects. In Table S1 of the present appendix, we therefore report the correlations between the neighborhood variables that were retained in the final model for study participation. These correlations were estimated at the level of the 2218 neighborhoods (except when there were missing values for some of the variables). We relied on the ordinal versions of the neighborhood variables to determine the correlations, as these ordinal variables were used in the modeling process. As shown in Table S1, the correlation between the variables was between –0.38 and +0.78.

To go further in the assessment of problems of multicollinearity,[19-20] we examined whether there were neighborhoods represented in each cell of the 4 × 4 cross-tabulations between neighborhood variables considered 2 by 2. In Tables S2-A, S2-B, and S2-C of the present appendix, we provide examples of cross-tabulations between the neighborhood variables (we selected 3 examples involving variables that were correlated with each other to a different extent).

In almost all cases, because of the large neighborhood sample size (n > 2200), there were neighborhoods represented in all cells of the cross-tabulations. However, in the most extreme case, the correlation was particularly high between neighborhood income and neighborhood dwelling value (r = 0.78). As shown in Table S2-C, there was no neighborhood with a high income that also had low dwelling values (there were on the opposite 20 neighborhoods with a low income and high dwelling values).

Therefore, the rate ratio for study participation for the 4th vs. the 1st neighborhood income quartiles cannot be estimated in the low neighborhood dwelling value stratum. However, rate ratios for the 2nd or 3rd vs. the 1st income quartiles can be estimated in this particular neighborhood dwelling value stratum, and rate ratios for the 4th vs. the 1st income quartiles can be estimated in the other neighborhood dwelling value strata.

Overall, we believe that the effects on study participation identified for neighborhood income and neighborhood dwelling value reflect connected but independent neighborhood influences that are separable to a certain extent. If only one of the two dimensions was truly associated with study participation, this particular variable would have captured all of the effect, which was not the case. Weighing the pros and cons, we found it more informative to maintain neighborhood income and dwelling value into the model, even if interpreting their effects requires caution.

D – Additional information on the RECORD Study Sample

D1 – Population recruited in the RECORD Cohort Study

In France, all working and retired employees and their families (either from French citizenship or not) are affiliated with the National Health Insurance System for Salaried Workers. As such, they are offered a free 2 hour long preventive medical examination every 5 years (people have to wait at least 5 years after their previous health checkup to benefit from another health checkup for free). A particular exception exists for people in situation of job insecurity or socioeconomic precariousness (unemployed persons, people with social allowances, homeless people, etc.), who have access to this free health checkup every year.

The following occupational categories are not affiliated with the French National Health Insurance System for Salaried Workers, and therefore could not be recruited in our study (they receive their health checkups in other health centers than those in which the RECORD participants were recruited):

- shopkeepers;
- craftsmen;
- farmers and salaried farm workers;
- the professions (lawyers, non-salaried physicians and healthcare professionals, architects, etc.).

However, in the Paris Ile-de-France region where the RECORD Cohort was recruited, working and retired employees and their families represent almost 95% of the population (i.e., about 10.26 millions of people out of 10.83 millions).

D2 – Distribution of study participants and total population according to individual and neighborhood characteristics

The distribution of study participants and total population according to individual and neighborhood characteristics is reported in Table S3 of the present appendix.

E – Bayesian modeling of individual/neighborhood effects on participation in the RECORD Study and on the odds of diabetes

E1 – Modeling strategy to construct the final model for study participation

The analytical strategy described in the main text of the article is a summary of the analyses that were performed. First we estimated separate models for study participation that were adjusted for age, gender, and individual education and each included a unique neighborhood variable. Neighborhood variables were then included two by two into the models, and a third variable, a fourth variable, etc., were progressively added. We checked to make sure there was no substantial modification of the associations when including an additional neighborhood variable into the model.

The only exception was the association between the proportion of people looking for work in the neighborhood and study participation, which changed from negative to positive after adjustment for neighborhood variables such as neighborhood income or neighborhood dwelling value. However, we deliberately decided to keep all these variables in the model because we had very definite hypotheses explaining why antagonistic effects on study participation were observed for neighborhood income or dwelling value on the one hand, and the proportion looking for work on the other hand (see the discussion section of the main article). Based on these hypotheses, it was logical to expect that a true positive effect of the proportion looking for work would be hidden by the effects of the other socioeconomic variables if not adjusted for them.

When all neighborhood variables independently associated with study participation were identified, we tested interactions between the effects of these variables and individual education.

E2 – Modeling strategy to construct the final model for diabetes

In the initial steps of the present study, three metabolic risk factors were considered: obesity, hypertension, and diabetes. There was indication that neighborhood effects on study participation only biased the association between neighborhood socioeconomic position and diabetes: residual geographic variations in study participation were not associated with obesity or hypertension, but were associated with diabetes. Therefore, the present article exclusively focuses on this metabolic outcome.

First we examined the individual sociodemographic correlates of diabetes. The following individual variables were included into the model: age and age square, gender, marital status, education, and perceived financial strain. We used a binary variable for marital status (living alone or cohabiting). Individual education was defined in the same way than in the analyses of study participation (no education; secondary school and lower tertiary education; and higher tertiary education). A binary variable for self-reported financial strain was determined.

Based on this model containing individual-level variables, we estimated 3 separate models to test the associations between neighborhood education, neighborhood income, or neighborhood dwelling value on the one hand, and the odds of diabetes on the other hand. There was a pattern of association between neighborhood income or neighborhood dwelling value and the odds of diabetes, but trend tests did not confirm these associations. Only neighborhood education was associated with diabetes.

E3 – General description of the models

In Boxes S1 to S3 of Online Appendix 2, we report the Winbugs code for some of the models that were estimated:

- Box S1 provides the code for the multilevel model for study participation that includes all neighborhood variables and an interaction between distance to the closest center and individual education (reported in Table 2 of the main article);
- Box S2 provides the code for the multilevel model for diabetes that includes the individual and neighborhood variables retained in the model and the median of the posterior distribution of each neighborhood's random effect for study participation as an explanatory variable (reported in Table 3, column 2, of the main article);
- Box S3 provides the code for the joint model for study participation and diabetes, in which the neighborhood random effect for study participation is injected as an explanatory variable in the model for diabetes.

In all these models, we have used a flat prior for the intercept, normal priors for the fixed effects, and gamma priors for random effect precisions.

E4 – General strategy to estimate the models

For each of the models, we ran 2 chains for 50000 iterations as a burn-in period, evaluated convergence of the chains to the posterior distribution of the parameters with the Gelman-Rubin-Brooks statistic, and checked that the Monte Carlo error for each parameter was lower than 5% of the standard deviation of the parameter. We then ran the two chains for 20000 additional iterations to derive the posterior distribution of the parameters. In Boxes S1 to S3 (Online Appendix 2), along with the Winbugs code, we provide the inits that were used for each of the two chains.

We present the median of the posterior distributions as parameter estimates, and use the 2.5^{th} and 97.5^{th} percentiles of the distributions as 95% credible intervals. The Markov chain Monte Carlo framework also allowed us to derive 95% credible intervals for the Moran's I coefficients determined from the neighborhood random effects of the multilevel models.

As an example, we have noted the run time of the model reported in Table 4 of the main text and Box S3 of Online Appendix 2 (which simultaneously estimates a regression equation for study participation and a regression equation for diabetes). This model was estimated with a computer with a 3.06 GHz Intel Xeon processor and 3.5 Go of RAM. Running the two chains of the model for 50000 iterations took 34.7 hours. After the model had converged and the Deviance Information Criterion (DIC) was required, it took 14.3 hours to run the two chains for 20000 additional iterations, from iteration 50000 to iteration 70000.

E5 – Comparison of different priors for the neighborhood random effect

Important geographic variations in the rate of study participation were identified in our analyses. Our expectation was that the hyperparameters of the gamma prior for the neighborhood random effect precision in the participation model do not have a significant influence on the *a posteriori* distribution of the random effect variance. However, we performed a sensitivity analysis in the choice of the hyperparameters of the gamma prior.

We report the results of the sensitivity analysis conducted for the multilevel model for study participation that only includes age and gender as explanatory variables. Three different priors were specified for the random effect precision, i.e., dgamma(0.5, 0.0005), dgamma(0.01, 0.01), and dgamma(0.001, 0.001). The conclusion of this sensitivity analysis is that the results were virtually identical in the three estimations. The random effect variance was 0.214 (95% credible interval: 0.183, 0.249) in the first case, 0.214 (95% credible interval: 0.183, 0.249) in the second case, and 0.214 (95% credible interval: 0.182, 0.248) when specifying the third gamma prior.

E6 – Comparison of our approach with Heckman selection models

In the present study, we performed a joint modeling of the neighborhood determinants of study participation and of the association between neighborhood socioeconomic status and diabetes. The neighborhood random effect of the model for study participation was incorporated as an explanatory variable in the model for diabetes for bias correction.

In its spirit, our approach is very close to the one implemented with Heckman selection models. In a seminal article,[21] James J. Heckman described biases resulting from using nonrandomly selected samples in regression analyses as a specification error that can be sometimes corrected by incorporating omitted variables as regressors into the model for the outcome of interest.

Heckman noted that in situations in which information is missing for the outcome for a number of observations (which observations cannot enter into the sample), the critical question is "why are the data missing?". He emphasized the importance of parameters of the function determining the probability of entrance into the sample.

He proposed to implement bias correction through the estimation of two different equations, the first one modeling the probability of entrance into the sample and the second one using this information for correction in the modeling of the outcome of interest. In the first stage, a regression is estimated for the likelihood of participating in the study. A selection bias parameter is generated that summarizes information about the factors that influence participation and consequently the observation of the outcome variable of interest. The selection bias parameter is included as an effect in the second stage model for the main outcome.

Obviously, all of these characteristics exactly apply to our proposed bias correction strategy. As a particular case of this general perspective, our approach is an original development that proposes to rely on the neighborhood random effect of a model for study participation to capture residual geographic variations in participation and adjust for their biasing effect.

F – Individual/area characteristics associated with attendance to the health center (in the general population) and with participation in the RECORD Study (among people visiting the health center)

Even though it was not necessary for our bias assessment perspective, a complementary analysis was conducted to distinguish between (i) contextual determinants of going for a health checkup (which is available for free to 95% of the population) and (ii) contextual determinants of inclusion in / exclusion from the study and acceptance / refusal to participate among people attending the centers. In this analysis, exclusion from the study, for individuals who were eligible based on age and residence, refers to their non-selection by the research staff because of their insufficient mastery of the French language or a cognitive limitation not allowing them to answer the questionnaire.

As shown in Table S4 of the present appendix, three distinct models were estimated to assess contextual determinants of (i) the rate of people participating in the study (in the general population), (ii) the rate of people going to the health centers for a checkup, whether participating or not (in the general population), and (iii) the likelihood of participating in the study (among individuals who came to the health centers for a checkup and who were eligible based on age and residence).

It should be noted, however, that eligible individuals who were excluded, refused to participate, or withdrew during the data collection process were not geocoded with as much precision as were subjects who were included in the study. Thus, the complementary analyses proposed here were conducted with individuals nested within 121 municipalities or large sections of Paris (rather than within neighborhoods), at which level all environmental variables were redefined.

As reported in Table S4, a Poisson model was used for outcomes (i) and (ii) listed above, and a logistic model was employed for outcome (iii). All models included a municipality unstructured random effect. After controlling for age, gender, and individual education, only area variables that were associated with the outcomes were retained.

As expected, we found that municipality-level variables were associated with the rate of people attending the health centers for a checkup, but not (or only marginally) with study participation among persons attending for the checkup. In coherence with the associations identified at the neighborhood level, a short distance to the center, a high area socioeconomic level, and a low density were associated with higher rates of people attending the health centers, even if slightly different variables were retained (i.e., municipality population density rather than building density and building height in the neighborhood-level model).

Strikingly, a model estimated among people attending the health centers indicated that, among contextual variables, only mean dwelling value was associated, and weakly so, with study participation. Conversely, in this model, individual education was strongly associated with participation, reflecting exclusion by the research staff of persons who were linguistically or cognitively unable to fill out study questionnaires, or low-educated persons refusing to participate or withdrawing during the data collection process.

References

1. Hernan MA, Hernandez-Diaz S, Robins JM. A structural approach to selection bias. *Epidemiology*. 2004;15:615-25.
2. Fleischer NL, Diez Roux AV. Using directed acyclic graphs to guide analyses of neighbourhood health effects: an introduction. *J Epidemiol Community Health*. 2008;62:842-6.
3. Greenland S. Quantifying biases in causal models: classical confounding vs collider-stratification bias. *Epidemiology*. 2003;14:300-6.
4. Chaix B, Leal C, Evans DW. Neighborhood-level confounding in epidemiologic studies: unavoidable challenges, uncertain solutions. *Epidemiology*. 2009; in press.
5. Greenland S. Response and follow-up bias in cohort studies. *Am J Epidemiol*. 1977;106:184-7.
6. Ferrie JE, Kivimaki M, Singh-Manoux A, et al. Non-response to baseline, non-response to follow-up and mortality in the Whitehall II cohort. *Int J Epidemiol*. 2009.
7. Cole SR, Platt RW, Schisterman EF, et al. Illustrating bias due to conditioning on a collider. *Int J Epidemiol*. 2010;39:417-20.
8. Diez Roux AV, Jacobs DR, Kiefe CI. Neighborhood characteristics and components of the insulin resistance syndrome in young adults: the coronary artery risk development in young adults (CARDIA) study. *Diabetes Care*. 2002;25:1976-82.
9. Auchincloss AH, Diez Roux AV, Brown DG, Erdmann CA, Bertoni AG. Neighborhood resources for physical activity and healthy foods and their association with insulin resistance. *Epidemiology*. 2008;19:146-57.
10. Auchincloss AH, Diez Roux AV, Brown DG, O'Meara ES, Raghunathan TE. Association of insulin resistance with distance to wealthy areas: the multi-ethnic study of atherosclerosis. *Am J Epidemiol*. 2007;165:389-97.
11. Schootman M, Andresen EM, Wolinsky FD, et al. The effect of adverse housing and neighborhood conditions on the development of diabetes mellitus among middle-aged African Americans. *Am J Epidemiol*. 2007;166:379-87.
12. Franco M, Diez Roux AV, Glass TA, Caballero B, Brancati FL. Neighborhood characteristics and availability of healthy foods in Baltimore. *Am J Prev Med*. 2008;35:561-7.
13. Diez Roux AV. Residential environments and cardiovascular risk. *J Urban Health*. 2003;80:569-589.
14. Moore LV, Diez Roux AV, Evenson KR, McGinn AP, Brines SJ. Availability of recreational resources in minority and low socioeconomic status areas. *Am J Prev Med*. 2008;34:16-22.
15. Raudenbush SW, Sampson RJ. Ecometrics: Toward a Science of Assessing Ecological Settings, With Application to the Systematic Social Observation of Neighborhoods. *Sociol Methodol*. 1999;29:1-41.
16. Mujahid MS, Diez Roux AV, Morenoff JD, Raghunathan T. Assessing the measurement properties of neighborhood scales: from psychometrics to ecometrics. *Am J Epidemiol*. 2007;165:858-67.
17. Chaix B, Lindstrom M, Merlo J, Rosvall M. Neighbourhood social interactions and risk of acute myocardial infarction. *J Epidemiol Community Health*. 2008;62:62-8.
18. Mujahid MS, Diez Roux AV, Borrell LN, Nieto FJ. Cross-sectional and longitudinal associations of BMI with socioeconomic characteristics. *Obes Res*. 2005;13:1412-1421.
19. Chaix B, Rosvall M, Lynch J, Merlo J. Disentangling contextual effects on cause-specific mortality in a longitudinal 23-year follow up study: impact of population density or socioeconomic environment? *Int J Epidemiol*. 2006;35:633-643.

20. Merlo J, Chaix B. Neighbourhood effects and the real world beyond randomized community trials: a reply to Michael J. Oakes. *Int J Epidemiol*. 2006;35:1361-1363.
21. Heckman JJ. Sample selection bias as a specification error. *Econometrica*. 1979;47:153-161.

TABLE S1. Pearson correlations (95% confidence intervals) between neighborhood factors expressed as ordinal variables retained in the final multilevel model for study participation reported in the main article (n = 2218 neighborhoods)

	Distance to the center	Median income	Proportion looking for work	Mean dwelling value	Proportion of the area covered by buildings	Mean building height
Distance to the center	1	–0.22 (–0.26, –0.18)	+0.01 (–0.03, +0.06)	–0.20 (–0.24, –0.16)	–0.38 (–0.41, –0.34)	–0.30 (–0.34, –0.26)
Median income		1	–0.25 (–0.29, –0.21)	+0.78 (+ 0.76, 0.80)	+0.13 (+0.09, +0.17)	+0.03 (–0.02, +0.07)
Proportion looking for work			1	–0.25 (–0.29, –0.21)	+0.02 (–0.03, +0.06)	+0.03 (–0.01, +0.07)
Mean dwelling value				1	+0.02 (–0.02, +0.07)	+0.07 (+0.03, +0.11)
Proportion of the area covered by buildings					1	+0.41 (+0.37, +0.44)
Mean buildings height						1

TABLE S2-A. Cross-tabulation between neighborhood median income and the proportion of the area covered by buildings in the neighborhood (weak correlation): number of neighborhoods in each cell (n = 2207 neighborhoods, 11 missing values)

Neighborhood median income / Proportion covered by buildings	First quartile	Second quartile	Third quartile	Fourth quartile
First quartile	197	141	121	132
Second quartile	184	159	113	108
Third quartile	132	138	136	125
Fourth quartile	90	117	155	159

TABLE S2-B. Cross-tabulation between the proportion of the area covered by buildings and mean building height in the neighborhood (intermediate level of correlation): number of neighborhoods in each cell (n = 2218)

Mean building height \ Proportion covered by buildings	First quartile	Second quartile	Third quartile	Fourth quartile
First quartile	288	207	76	13
Second quartile	153	207	162	59
Third quartile	97	88	144	210
Fourth quartile	64	62	149	239

TABLE S2-C. Cross-tabulation between neighborhood median income and neighborhood dwelling value (strong correlation): number of neighborhoods in each cell (n = 2207 neighborhoods, 11 missing values)

Neighborhood median income / Neighborhood dwelling value	First quartile	Second quartile	Third quartile	Fourth quartile
First quartile	401	181	14	0
Second quartile	134	230	159	18
Third quartile	48	105	233	135
Fourth quartile	20	39	119	371

TABLE S3. Distribution of study participants (RECORD Study) and total population (Population Census) according to individual and neighborhood characteristics

	RECORD sample	Total population
Age		
30–39 years	21.3%	28.3%
40–59 years	55.8%	43.5%
60 years and over	22.9%	28.3%
Men	65.6%	46.8%
Individual education level		
Low	7.8%	17.6%
Medium	54.1%	64.1%
High	38.0%	18.3%
Distance to the center		
Low	31.8%	25.0%
Mid-low	26.1%	25.0%
Mid-high	22.5%	25.0%
High	19.5%	25.0%
Median income		
Low	17.9%	25.0%
Mid-low	22.2%	25.0%
Mid-high	26.2%	25.0%
High	33.7%	25.0%
Mean dwelling value		
Low	18.5%	25.0%
Mid-low	22.6%	25.0%
Mid-high	25.9%	25.0%
High	33.1%	25.0%
Proportion of the active population looking for work		
Low	31.0%	25.0%
Mid-low	25.5%	25.0%
Mid-high	23.0%	25.0%
High	20.5%	25.0%
Proportion of the area covered by buildings		
Low	26.1%	25.0%
Mid-low	24.5%	25.0%
Mid-high	24.8%	25.0%
High	24.6%	25.0%
Mean building height		
Low	25.9%	25.0%
Mid-low	24.5%	25.0%
Mid-high	24.7%	25.0%
High	24.9%	25.0%

TABLE S4. Effects of municipality environmental variables on (i) the rate of participation in the RECORD Cohort Study (in the general population), (ii) the rate of people attending the health centers (in the general population), and (iii) the odds of participation in the study (among people attending the health centers), as estimated from multilevel regression models adjusted for individual characteristics (all effects adjusted for each other)

	Outcome : participation in the study (general population)		**Outcome: attending the health center (general population)**		**Outcome: participation in the study (people attending the health centers)**	
	RR[a]	**95% CI**	**RR[a]**	**95% CI**	**OR[a,b]**	**95% CI**
Age (vs. 30–39 years)						
40–59 years	1.86	(1.76, 1.98)	1.91	(1.82, 2.00)	0.93	(0.84, 1.03)
60 years and over	1.38	(1.29, 1.48)	1.34	(1.27, 1.42)	0.95	(0.84, 1.08)
Men (vs. women)	1.99	(1.90, 2.10)	1.90	(1.83, 1.97)	1.07	(0.99, 1.17)
Individual education (vs. low)						
Medium	1.89	(1.74, 2.07)	0.76	(0.72, 0.80)	5.51	(4.92, 6.18)
High	4.26	(3.89, 4.68)	1.59	(1.51, 1.69)	7.24	(6.37, 8.25)
Distance to the center (vs. high)						
Mid-high	1.10	(0.93, 1.30)	1.14	(0.97, 1.34)		-
Mid-low	1.44	(1.20, 1.72)	1.53	(1.27, 1.84)		-
Low	1.62	(1.34, 1.94)	1.90	(1.57, 2.28)		-
Mean dwelling value (vs. low)						
Mid-low	1.26	(1.05, 1.50)	1.18	(0.98, 1.41)	1.19	(1.01, 1.41)
Mid-high	1.28	(1.09, 1.51)	1.21	(1.03, 1.43)	1.11	(0.94, 1.30)
High	1.44	(1.22, 1.71)	1.34	(1.13, 1.59)	1.20	(1.02, 1.41)
Population density (vs. high)						
Mid-high	1.16	(0.90, 1.49)	1.17	(0.92, 1.51)		-
Mid-low	1.28	(0.98, 1.64)	1.36	(1.07, 1.76)		-
Low	1.57	(1.22, 1.98)	1.72	(1.37, 2.19)		-

[a]RR, rate ratio; OR, odds ratio.

[b]Missing information in the third column of the Table indicates that the corresponding variables were not associated with study participation.

Figure S1-A. Influences of neighborhood socioeconomic position and obesity on study participation as a source of collider bias. The dashed line represents the association generated by restricting the analyses to participants. Based on Hernán (Epidemiology 2004;15:615-25), the rectangle around participation indicates that the analyses condition on participation.

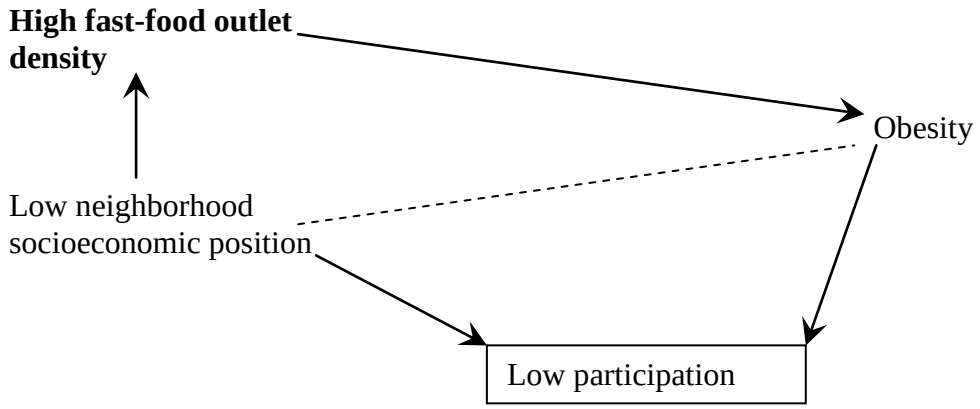
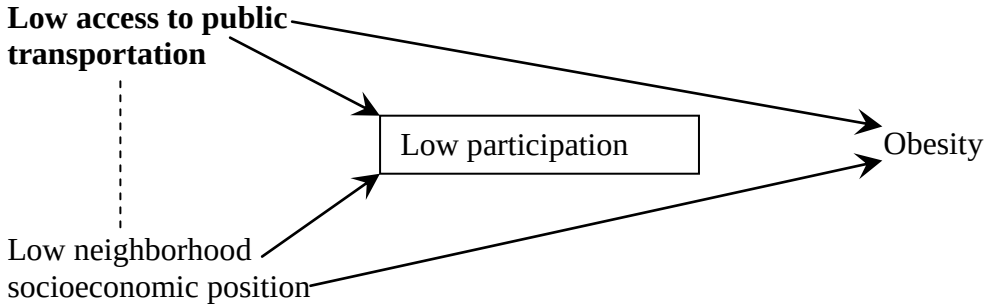


Figure S1-B. Influences of neighborhood socioeconomic position and access to public transportation on study participation as a source of collider bias. The dashed line represents the association generated by restricting the analyses to participants. Based on Hernán (Epidemiology 2004;15:615-25), the rectangle around participation indicates that the analyses condition on participation.



A joint modeling of neighborhood effects on participation in the RECORD Cohort Study and neighborhood effects on type 2 diabetes: bias assessment and correction

Online Appendix 2

BOX S1. Winbugs code for the multilevel model for study participation that includes all neighborhood variables and an interaction between distance to the closest center and individual education

```
model
{
        for (i in 1 : 39127)
                {
                ptcp_profil[i]~dpois(mu[i])

log(mu[i])<- log(N_profil[i]) + alpha + unstr[codiris[i]] + beta1[homme[i]] + beta2[age_2[i]] +
beta3[age_3[i]] + beta6[rvmeduc20054_2[i]] + beta7[rvmeduc20054_3[i]] + beta8[rvmeduc20054_4[i]]
+ beta9[rvmeduc20054_m[i]] + beta14[ppt_dem_emploi_20064_2[i]] +
beta15[ppt_dem_emploi_20064_3[i]] + beta16[ppt_dem_emploi_20064_4[i]] +
beta17[ppt_dem_emploi_20064_m[i]] + beta18[housingrank_iris4_2[i]]  +
beta19[housingrank_iris4_3[i]] + beta20[housingrank_iris4_4[i]] + beta24[pbuilt_surface_iris4_3[i]] +
beta25[pbuilt_surface_iris4_2[i]] +beta26[pbuilt_surface_iris4_1[i]]  + beta27[mbuild_height_iris4_3[i]]
+beta28[mbuild_height_iris4_2[i]] +beta29[mbuild_height_iris4_1[i]]

+ beta30[inter_dist_etude_2[i]] + beta31[inter_dist_etude_3[i]] + beta32[inter_dist_etude_4[i]]
+ beta33[inter_dist_etude_5[i]] + beta34[inter_dist_etude_6[i]] + beta35[inter_dist_etude_7[i]]
+ beta36[inter_dist_etude_8[i]] + beta37[inter_dist_etude_9[i]] + beta38[inter_dist_etude_10[i]]
+ beta39[inter_dist_etude_11[i]] + beta40[inter_dist_etude_12[i]]
                }

for (j in 1 : 2218)
        { unstr[j]~dnorm(m.unstr, tau.unstr) }

alpha~dflat()

beta1[1]<-0
beta1[2] ~dnorm(0,0.00001)

beta2[1]<-0
beta2[2] ~dnorm(0,0.00001)

beta3[1]<-0
beta3[2] ~dnorm(0,0.00001)

beta6[1]<-0
beta6[2] ~ dnorm(0,0.00001)

beta7[1]<-0
beta7[2] ~ dnorm(0,0.00001)

beta8[1]<-0
beta8[2] ~ dnorm(0,0.00001)

beta9[1]<-0
beta9[2] ~ dnorm(0,0.00001)

beta14[1]<-0
beta14[2] ~ dnorm(0,0.00001)

beta15[1]<-0
beta15[2] ~ dnorm(0,0.00001)
```

```
beta16[1]<-0
beta16[2] ~ dnorm(0,0.00001)

beta17[1]<-0
beta17[2] ~ dnorm(0,0.00001)

beta18[1]<-0
beta18[2] ~ dnorm(0,0.00001)

beta19[1]<-0
beta19[2] ~ dnorm(0,0.00001)

beta20[1]<-0
beta20[2] ~ dnorm(0,0.00001)

beta24[1]<-0
beta24[2] ~ dnorm(0,0.00001)

beta25[1]<-0
beta25[2] ~ dnorm(0,0.00001)

beta26[1]<-0
beta26[2] ~ dnorm(0,0.00001)

beta27[1]<-0
beta27[2] ~ dnorm(0,0.00001)

beta28[1]<-0
beta28[2] ~ dnorm(0,0.00001)

beta29[1]<-0
beta29[2] ~ dnorm(0,0.00001)

beta30[1]<-0
beta30[2] ~ dnorm(0,0.00001)

beta31[1]<-0
beta31[2] ~ dnorm(0,0.00001)

beta32[1]<-0
beta32[2] ~ dnorm(0,0.00001)

beta33[1]<-0
beta33[2] ~ dnorm(0,0.00001)

beta34[1]<-0
beta34[2] ~ dnorm(0,0.00001)

beta35[1]<-0
beta35[2] ~ dnorm(0,0.00001)

beta36[1]<-0
beta36[2] ~ dnorm(0,0.00001)

beta37[1]<-0
beta37[2] ~ dnorm(0,0.00001)

beta38[1]<-0
beta38[2] ~ dnorm(0,0.00001)
```

```
beta39[1]<-0
beta39[2] ~ dnorm(0,0.00001)

beta40[1]<-0
beta40[2] ~ dnorm(0,0.00001)

m.unstr<-0
tau.unstr~dgamma(0.5,0.0005)
var.unstr<-1/tau.unstr

#Empirical marginal variances

sdme.unstr<-sd(unstr[])
varme.unstr<-pow(sdme.unstr,2)
}
```

INITS CHAIN 1

```
list(alpha=0, tau.unstr=1, beta1=c(NA,0),  beta2=c(NA,0), beta3=c(NA,0), beta6=c(NA,0),
beta7=c(NA,0),beta8=c(NA,0), beta9=c(NA,0), beta14=c(NA,0), beta15=c(NA,0), beta16=c(NA,0),
beta17=c(NA,0), beta18=c(NA,0),  beta19=c(NA,0), beta20=c(NA,0), beta24=c(NA,0),
beta25=c(NA,0), beta26=c(NA,0), beta27=c(NA,0), beta28=c(NA,0), beta29=c(NA,0), beta30=c(NA,0),
beta31=c(NA,0), beta32=c(NA,0), beta33=c(NA,0), beta34=c(NA,0), beta35=c(NA,0), beta36=c(NA,0),
beta37=c(NA,0), beta38=c(NA,0), beta39=c(NA,0), beta40=c(NA,0), unstr=c(0,0,0, …, 0,0,0))
```

INITS CHAIN 2

```
list(alpha=0.5,tau.unstr=10, beta1=c(NA,0.5), beta2=c(NA,0.5), beta3=c(NA,0.5), beta6=c(NA,0.5),
beta7=c(NA,0.5),beta8=c(NA,0.5), beta9=c(NA,0.5), beta14=c(NA,0.5), beta15=c(NA,0.5),
beta16=c(NA,0.5), beta17=c(NA,0.5), beta18=c(NA,0.5),  beta19=c(NA,0.5), beta20=c(NA,0.5),
beta24=c(NA,0.5), beta25=c(NA,0.5), beta26=c(NA,0.5), beta27=c(NA,0.5), beta28=c(NA,0.5),
beta29=c(NA,0.5), beta30=c(NA,0.5), beta31=c(NA,0.5), beta32=c(NA,0.5), beta33=c(NA,0.5),
beta34=c(NA,0.5), beta35=c(NA,0.5), beta36=c(NA,0.5), beta37=c(NA,0.5), beta38=c(NA,0.5),
beta39=c(NA,0.5), beta40=c(NA,0.5), unstr=c(0,0,0, …, 0,0,0))
```

BOX S2. Winbugs code for the multilevel model for diabetes that includes the individual and neighborhood variables retained in the model and the median of the posterior distribution of each neighborhood's random effect for study participation as an explanatory variable divided into four categories

```
model
{
  for(i in 1 : 6876) {
     diabete[i] ~ dbern(p[i])
     logit(p[i]) <- alpha + beta1*age[i] + beta2*agecarre[i] + beta3[homme[i]] + beta4[cohab_seul[i]] +
beta5[financial_strain[i]] + beta6[financial_strain_miss[i]] + beta7[nivetude_2[i]] + beta8[nivetude_1[i]]
+ beta9[nivetude_m[i]] + beta10[peducsup_iris_4_3[i]] + beta11[peducsup_iris_4_2[i]] +
beta12[peducsup_iris_4_1[i]] + beta13[participation_2[i]] + beta14[participation_3[i]] +
beta15[participation_4[i]] + alea[codiris_indiv[i]]
  }

alpha ~ dflat()

beta1 ~ dflat()

beta2 ~ dflat()

beta3[1]<-0
beta3[2] ~ dnorm(0,0.00001)

beta4[1]<-0
beta4[2] ~ dnorm(0,0.00001)

beta5[1]<-0
beta5[2] ~ dnorm(0,0.00001)

beta6[1]<-0
beta6[2] ~ dnorm(0,0.00001)

beta7[1]<-0
beta7[2] ~ dnorm(0,0.00001)

beta8[1]<-0
beta8[2] ~ dnorm(0,0.00001)

beta9[1]<-0
beta9[2] ~ dnorm(0,0.00001)

beta10[1]<-0
beta10[2] ~ dnorm(0,0.00001)

beta11[1]<-0
beta11[2] ~ dnorm(0,0.00001)

beta12[1]<-0
beta12[2] ~ dnorm(0,0.00001)

beta13[1]<-0
beta13[2] ~ dnorm(0,0.00001)

beta14[1]<-0
beta14[2] ~ dnorm(0,0.00001)

beta15[1]<-0
```

```
beta15[2] ~ dnorm(0,0.00001)

for (j in 1 : 1882) {alea[j] ~ dnorm(Malea,Tau.alea)}
Malea <- 0
Tau.alea ~ dgamma(0.5,0.0005)
Varalea <- 1/Tau.alea
}
```

INITS CHAIN 1

```
list(alpha=-12, Tau.alea=1, beta1=0, beta2=0, beta3=c(NA,0), beta4=c(NA,0), beta5=c(NA,0),
beta6=c(NA,0), beta7=c(NA,0), beta8=c(NA,0), beta9=c(NA,0), beta10=c(NA,0), beta11=c(NA,0),
beta12=c(NA,0), beta13=c(NA,0), beta14=c(NA,0), beta15=c(NA,0), alea=c(0,0,0, …, 0,0,0))
```

INITS CHAIN 2

```
list(alpha=-12, Tau.alea=10, beta1=0.5, beta2=0, beta3=c(NA,0.5), beta4=c(NA,0.5), beta5=c(NA,0.5),
beta6=c(NA,0.5), beta7=c(NA,0.5), beta8=c(NA,0.5), beta9=c(NA,0.5), beta10=c(NA,0.5),
beta11=c(NA,0.5), beta12=c(NA,0.5), beta13=c(NA,0), beta14=c(NA,0), beta15=c(NA,0),
alea=c(0,0,0, …, 0,0,0))
```

BOX S3. Winbugs code for the joint model for study participation and diabetes: the neighborhood random effect for study participation is inserted as an explanatory variable in the model for diabetes

```
model
{
#Model for participation

for (i in 1 : 39127) {
   ptcp_profil[i]~dpois(mu[i])
log(mu[i])<- log(N_profil[i]) + palpha + unstr[codiris[i]] + pbeta1[homme_rp[i]] + pbeta2[age_rp_2[i]] +
pbeta3[age_rp_3[i]] + pbeta6[rvmeduc20054_2[i]] + pbeta7[rvmeduc20054_3[i]] +
pbeta8[rvmeduc20054_4[i]]  + pbeta9[rvmeduc20054_m[i]] + pbeta14[ppt_dem_emploi_20064_2[i]] +
pbeta15[ppt_dem_emploi_20064_3[i]] + pbeta16[ppt_dem_emploi_20064_4[i]] +
pbeta17[ppt_dem_emploi_20064_m[i]] + pbeta18[housingrank_iris4_2[i]]  +
pbeta19[housingrank_iris4_3[i]] + pbeta20[housingrank_iris4_4[i]] + pbeta24[pbuilt_surface_iris4_3[i]]
+ pbeta25[pbuilt_surface_iris4_2[i]] +pbeta26[pbuilt_surface_iris4_1[i]]  +
pbeta27[mbuild_height_iris4_3[i]] + pbeta28[mbuild_height_iris4_2[i]]
+pbeta29[mbuild_height_iris4_1[i]]

+ pbeta30[inter_dist_etude_2[i]] + pbeta31[inter_dist_etude_3[i]] + pbeta32[inter_dist_etude_4[i]]
+ pbeta33[inter_dist_etude_5[i]] + pbeta34[inter_dist_etude_6[i]] + pbeta35[inter_dist_etude_7[i]]
+ pbeta36[inter_dist_etude_8[i]] + pbeta37[inter_dist_etude_9[i]] + pbeta38[inter_dist_etude_10[i]]
+ pbeta39[inter_dist_etude_11[i]] + pbeta40[inter_dist_etude_12[i]]
                    }

for (j in 1 : 2218)
          { unstr[j]~dnorm(m.unstr, tau.unstr) }

palpha~dflat()

pbeta1[1]<-0
pbeta1[2] ~dnorm(0,0.00001)

pbeta2[1]<-0
pbeta2[2] ~dnorm(0,0.00001)

pbeta3[1]<-0
pbeta3[2] ~dnorm(0,0.00001)

pbeta6[1]<-0
pbeta6[2] ~ dnorm(0,0.00001)

pbeta7[1]<-0
pbeta7[2] ~ dnorm(0,0.00001)

pbeta8[1]<-0
pbeta8[2] ~ dnorm(0,0.00001)

pbeta9[1]<-0
pbeta9[2] ~ dnorm(0,0.00001)

pbeta14[1]<-0
pbeta14[2] ~ dnorm(0,0.00001)

pbeta15[1]<-0
pbeta15[2] ~ dnorm(0,0.00001)

pbeta16[1]<-0
```

```
pbeta16[2] ~ dnorm(0,0.00001)

pbeta17[1]<-0
pbeta17[2] ~ dnorm(0,0.00001)

pbeta18[1]<-0
pbeta18[2] ~ dnorm(0,0.00001)

pbeta19[1]<-0
pbeta19[2] ~ dnorm(0,0.00001)

pbeta20[1]<-0
pbeta20[2] ~ dnorm(0,0.00001)

pbeta24[1]<-0
pbeta24[2] ~ dnorm(0,0.00001)

pbeta25[1]<-0
pbeta25[2] ~ dnorm(0,0.00001)

pbeta26[1]<-0
pbeta26[2] ~ dnorm(0,0.00001)

pbeta27[1]<-0
pbeta27[2] ~ dnorm(0,0.00001)

pbeta28[1]<-0
pbeta28[2] ~ dnorm(0,0.00001)

pbeta29[1]<-0
pbeta29[2] ~ dnorm(0,0.00001)

pbeta30[1]<-0
pbeta30[2] ~ dnorm(0,0.00001)

pbeta31[1]<-0
pbeta31[2] ~ dnorm(0,0.00001)

pbeta32[1]<-0
pbeta32[2] ~ dnorm(0,0.00001)

pbeta33[1]<-0
pbeta33[2] ~ dnorm(0,0.00001)

pbeta34[1]<-0
pbeta34[2] ~ dnorm(0,0.00001)

pbeta35[1]<-0
pbeta35[2] ~ dnorm(0,0.00001)

pbeta36[1]<-0
pbeta36[2] ~ dnorm(0,0.00001)

pbeta37[1]<-0
pbeta37[2] ~ dnorm(0,0.00001)

pbeta38[1]<-0
pbeta38[2] ~ dnorm(0,0.00001)

pbeta39[1]<-0
```

```
pbeta39[2] ~ dnorm(0,0.00001)

pbeta40[1]<-0
pbeta40[2] ~ dnorm(0,0.00001)

m.unstr<-0
tau.unstr~dgamma(0.5,0.0005)
var.unstr<-1/tau.unstr

#Empirical marginal variances

sdme.unstr<-sd(unstr[])
varme.unstr<-pow(sdme.unstr,2)

#Model for diabetes

for(i in 1 : 6876) {
    diabete[i] ~ dbern(p[i])
    logit(p[i]) <- alpha + beta1*age[i] + beta2*agecarre[i] + beta3[homme[i]] + beta4[cohab_seul[i]] +
beta5[financial_strain[i]] + beta6[financial_strain_miss[i]] + beta7[nivetude_2[i]] + beta8[nivetude_1[i]]
+ beta9[nivetude_m[i]] + beta10[peducsup_iris_4_3[i]] + beta11[peducsup_iris_4_2[i]] +
beta12[peducsup_iris_4_1[i]] + beta13*unstr[codiris_ptcp[i]] + alea[codiris_indiv[i]]
  }

alpha ~ dflat()

beta1 ~ dflat()

beta2 ~ dflat()

beta3[1]<-0
beta3[2] ~ dnorm(0,0.00001)

beta4[1]<-0
beta4[2] ~ dnorm(0,0.00001)

beta5[1]<-0
beta5[2] ~ dnorm(0,0.00001)

beta6[1]<-0
beta6[2] ~ dnorm(0,0.00001)

beta7[1]<-0
beta7[2] ~ dnorm(0,0.00001)

beta8[1]<-0
beta8[2] ~ dnorm(0,0.00001)

beta9[1]<-0
beta9[2] ~ dnorm(0,0.00001)

beta10[1]<-0
beta10[2] ~ dnorm(0,0.00001)

beta11[1]<-0
beta11[2] ~ dnorm(0,0.00001)

beta12[1]<-0
beta12[2] ~ dnorm(0,0.00001)
```

```
beta13 ~ dflat()

for (j in 1 : 1882) {alea[j] ~ dnorm(Malea,Tau.alea)}
Malea <- 0
Tau.alea ~ dgamma(0.5,0.0005)
Varalea <- 1/Tau.alea

}
```

INITS CHAINE 1

```
list(palpha=0,tau.unstr=1, pbeta1=c(NA,0),  pbeta2=c(NA,0), pbeta3=c(NA,0), pbeta6=c(NA,0),
pbeta7=c(NA,0), pbeta8=c(NA,0), pbeta9=c(NA,0), pbeta14=c(NA,0), pbeta15=c(NA,0),
pbeta16=c(NA,0), pbeta17=c(NA,0), pbeta18=c(NA,0),  pbeta19=c(NA,0), pbeta20=c(NA,0),
pbeta24=c(NA,0), pbeta25=c(NA,0), pbeta26=c(NA,0), pbeta27=c(NA,0), pbeta28=c(NA,0),
pbeta29=c(NA,0), pbeta30=c(NA,0), pbeta31=c(NA,0), pbeta32=c(NA,0), pbeta33=c(NA,0),
pbeta34=c(NA,0), pbeta35=c(NA,0), pbeta36=c(NA,0), pbeta37=c(NA,0), pbeta38=c(NA,0),
pbeta39=c(NA,0), pbeta40=c(NA,0), alpha=-12, Tau.alea=1, beta1=0, beta2=0, beta3=c(NA,0),
beta4=c(NA,0), beta5=c(NA,0), beta6=c(NA,0), beta7=c(NA,0), beta8=c(NA,0), beta9=c(NA,0),
beta10=c(NA,0), beta11=c(NA,0), beta12=c(NA,0), beta13=0, alea=c(0,0,0, …, 0,0,0))
```

INITS CHAINE 2

```
list(palpha=0.5,tau.unstr=10, pbeta1=c(NA,0.5),  pbeta2=c(NA,0.5), pbeta3=c(NA,0.5),
pbeta6=c(NA,0.5), pbeta7=c(NA,0.5), pbeta8=c(NA,0.5), pbeta9=c(NA,0.5), pbeta14=c(NA,0.5),
pbeta15=c(NA,0.5), pbeta16=c(NA,0.5), pbeta17=c(NA,0.5), pbeta18=c(NA,0.5), pbeta19=c(NA,0.5),
pbeta20=c(NA,0.5), pbeta24=c(NA,0.5), pbeta25=c(NA,0.5), pbeta26=c(NA,0.5), pbeta27=c(NA,0.5),
pbeta28=c(NA,0.5), pbeta29=c(NA,0.5), pbeta30=c(NA,0.5), pbeta31=c(NA,0.5), pbeta32=c(NA,0.5),
pbeta33=c(NA,0.5), pbeta34=c(NA,0.5), pbeta35=c(NA,0.5), pbeta36=c(NA,0.5), pbeta37=c(NA,0.5),
pbeta38=c(NA,0.5), pbeta39=c(NA,0.5), pbeta40=c(NA,0.5), alpha=-12, Tau.alea=1, beta1=0.5,
beta2=0, beta3=c(NA,0.5), beta4=c(NA,0.5), beta5=c(NA,0.5), beta6=c(NA,0.5), beta7=c(NA,0.5),
beta8=c(NA,0.5), beta9=c(NA,0.5), beta10=c(NA,0.5), beta11=c(NA,0.5), beta12=c(NA,0.5),
beta13=0, alea=c(0,0,0, …, 0,0,0))
```