

Practice of Epidemiology

A Simple Unified Approach for Estimating Natural Direct and Indirect Effects

Theis Lange*, Stijn Vansteelandt, and Maarten Bekaert

* Correspondence to Dr. Theis Lange, Department of Biostatistics, University of Copenhagen, Øster Farimagsgade 5, DK-1014 Copenhagen K, Denmark (e-mail: t.lange@biostat.ku.dk).

Initially submitted August 25, 2011; accepted for publication December 16, 2011.

An important problem within both epidemiology and many social sciences is to break down the effect of a given treatment into different causal pathways and to quantify the importance of each pathway. Formal mediation analysis based on counterfactuals is a key tool when addressing this problem. During the last decade, the theoretical framework for mediation analysis has been greatly extended to enable the use of arbitrary statistical models for outcome and mediator. However, the researcher attempting to use these techniques in practice will often find implementation a daunting task, as it tends to require special statistical programming. In this paper, the authors introduce a simple procedure based on marginal structural models that directly parameterize the natural direct and indirect effects of interest. It tends to produce more parsimonious results than current techniques, greatly simplifies testing for the presence of a direct or an indirect effect, and has the advantage that it can be conducted in standard software. However, its simplicity comes at the price of relying on correct specification of models for the distribution of mediator (and exposure) and accepting some loss of precision compared with more complex methods. Web Appendixes 1 and 2, which are posted on the *Journal's* Web site (<http://aje.oupjournals.org/>), contain implementation examples in SAS software (SAS Institute, Inc., Cary, North Carolina) and R language (R Foundation for Statistical Computing, Vienna, Austria).

causal inference; marginal structural models; mediation

Abbreviation: MSM, marginal structural model.

Important questions within both epidemiology and social sciences often require moving beyond "simply" estimating the total effect of a given exposure and instead require breaking down the total effect into separate causal pathways. The foremost example of such a strategy, which we consider here, is the decomposition of total effect into an indirect effect mediated through a specific mediator and the remaining direct effect.

The standard approach, which is inspired by Baron and Kenny (1), involves estimating the direct effect as the residual association between outcome and exposure after regression adjustment for the mediator(s) and the indirect effect by subtracting this from the total effect (on an appropriate scale). It has been shown that this approach works in the special case of linear models without interactions but is fundamentally flawed otherwise (2–4). Building on the counterfactual framework (refer to the work by Pearl (5)), a

formal approach to mediation analysis has now been developed. Using ideas of Robins and Greenland (6), Pearl (7) showed that a total effect can always be broken down into a so-called natural direct and indirect effect, regardless of the underlying statistical model.

Although much attention has been given to the development of identification conditions for natural direct and indirect effects (6–11), the researcher who wishes to estimate natural direct and indirect effects from actual data continues to face many challenges. This is because current procedures obtain natural direct and indirect effect estimates through a nontrivial combination of parameter estimates from a regression model for the mediator and a regression model for the outcome (4, 8, 10–15). The way to compute natural direct and indirect effects can therefore differ substantially between different types of mediator or outcome, and standard error calculations become even more tedious.

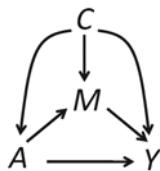


Figure 1. Directed acyclic graph of the causal structure assumed throughout the paper. Note that A is the exposure of interest; M , the mediator; C , a set of baseline confounders; and Y , the outcome.

Furthermore, even simple models for the mediator and outcome (e.g., a linear model for the mediator and a logistic regression model for the outcome) tend to produce complex expressions of natural direct and indirect effects. This can make results difficult to report (e.g., because these effects may turn out to depend on covariates in a complicated way). Moreover, it makes interesting hypotheses (e.g., for modification of the direct or indirect effect by covariates) essentially impossible to test, because it can be difficult to specify models for the outcome and mediator that satisfy the considered null hypothesis (e.g., that the natural (in)direct effect does not depend on covariates). All in all, this complexity places severe restrictions on the practical utility of current methods for computing natural direct and indirect effects for general epidemiologic research.

In this paper, we suggest a unified model for the direct and indirect effects with a corresponding simple estimation procedure. The estimation procedure is related to work by Hong (16) but generalized to a broad class of outcomes. The approach can, in principle, be used for any type of outcome (binary, continuous, survival, categorical, and so on) and any type of mediator, even when exposure-mediator interactions exist. Because it involves directly modeling the natural direct and indirect effects of interest, results become simpler for reporting, and interesting hypotheses concerning these effects become straightforward to test. The approach can be implemented in any software package capable of handling weighted modeling. However, its simplicity comes at the price of relying on correct specification of models for the distribution of mediator (and exposure) and of not exploiting all information in the data; more efficient estimators under the same model can thus be obtained. Web Appendixes 1 and 2, posted on the *Journal's* Web site (<http://aje.oupjournals.org/>), present detailed implementation examples in SAS software and R language.

DEFINITIONS AND ASSUMPTIONS

The results of this paper are based on the directed acyclic graph depicted in Figure 1, where A is the observed exposure of interest; M , the mediator; C , a set of baseline confounders; and Y , the outcome. Thus, it is assumed that there are no unmeasured confounders (confounders not included in C) for the exposure-outcome, exposure-mediator, or mediator-outcome relations. Variables are allowed to be of any type (e.g., continuous, binary, categorical, or survival). For each subject, we define the counterfactual variable

$Y_{a,m}$ as the outcome we would, possibly contrary to the fact, have observed for that subject had the exposure A been set to the value a and the mediator M set to m . Similarly, the counterfactual variable M_a denotes the value of the mediator if, possibly contrary to the fact, the exposure A was set to a .

Following the tradition in the causal inference literature (9), we will describe direct and indirect effects in terms of so-called nested counterfactuals, Y_{a^*,M_a} , denoting the outcome that would have been observed if A were set to a^* and M were set to the value it would have taken if A were set to a . In particular, we will compare Y_{a,M_a} with Y_{a^*,M_a} to obtain a measure of the natural direct effect of changing the exposure from a to a^* . Such comparison can, for instance, be made in terms of an average difference within levels of covariates, $E[Y_{a,M_a} - Y_{a^*,M_a}]|C$, or marginally, $E[Y_{a,M_a} - Y_{a^*,M_a}]$; as a risk ratio, $P[Y_{a,M_a} = 1]/P[Y_{a^*,M_a} = 1]$, and so on. Likewise, we will compare Y_{a^*,M_a} with $Y_{a^*,M_{a^*}}$ to obtain a measure of the natural indirect effect. The word "natural" refers to the fact that we let the mediator take the value it would take naturally when the exposure is set to a .

In this paper, as in most of the work on causal mediation analysis, we will discuss the estimation of natural direct and indirect effects under the no-unmeasured confounding assumption, implicit in the causal diagram of Figure 1, that the same set of covariates C is sufficient to control for confounding of the associations between exposure and outcome, exposure and mediator, and mediator and outcome. In particular, we thus assume that there are no variables L that are effects of exposure and that confound the mediator-outcome relation. A formal description of these assumptions is, for instance, given in the report by VanderWeele and Vansteelandt (4).

COUNTERFACTUAL-BASED MEDIATION ANALYSIS

The traditional approach to estimating natural direct and indirect effects uses the mediation formula (17) to calculate $E[Y_{a^*,M_a}|C = c]$ as

$$\sum_m E[Y|A = a^*, M = m, C = c] P(M = m|A = a, C = c). \quad (1)$$

This corresponds to estimating the mean value of the outcome in each stratum defined by mediator and confounders among the individuals with treatment a^* but weighting these by the likelihood of each mediator value among individuals with treatment a . Likewise, $E[Y_{a^*,M_a}]$ can be calculated as

$$\sum_c \sum_m E[Y|A = a^*, M = m, C = c] P(M = m|A = a, C = c) P(C = c). \quad (2)$$

When C is high dimensional, it will be necessary to use parametric models for the outcome mean and mediator distribution. If the outcome Y and mediator M are modeled by a linear model, that is, $E[Y|A = a^*, M = m, Z = z] = \alpha_0 + \alpha_1^* + \alpha_2 m + \alpha_3 z$ and $E[M|A = a, C = c] = \beta_0 + \beta_1 a + \beta_2 c$,

equation 2 simplifies greatly, and the natural direct and indirect effects are captured by $\alpha_1(a - a^*)$ and $\alpha_2\beta_1(a - a^*)$, respectively. However, equation 2 is less suited to outcomes modeled by a nonlinear model, for example, binary outcomes modeled by a logistic regression or survival outcomes, because the resulting expressions for the natural direct and indirect effects easily become complicated (e.g., they may depend on the values of the confounders in a complicated way; refer to the articles by VanderWeele and Vansteelandt (13) and by VanderWeele (15)).

ESTIMATING NATURAL EFFECTS BY MARGINAL STRUCTURAL MODELS

Marginal structural models (MSMs) are models for the marginal expectation (or distribution) of a counterfactual outcome (18). They have become popular for nonnested counterfactuals such as Y_a . For instance, the total causal effect of the exposure A on the outcome Y can be modeled in terms of a MSM of the form $E[Y_a] = b_0 + b_1a$, where b_1 then captures the average causal effect of the exposure. In contrast, MSMs for nested counterfactuals such as $Y_{a,M_{a^*}}$ have received very little attention, with few exceptions (19, 20). Such models are nonetheless of interest as they enable simultaneous and parsimonious modeling of the natural direct and indirect effect of the exposure A on the outcome Y other than through mediator M , as, for example,

$$E[Y_{a,M_{a^*}}] = c_0 + c_1a + c_2a^*. \quad (3)$$

Here, we have included the exposure twice to ascertain that it works through 2 distinct causal pathways. It is now easy to infer that $c_1(a - a^*)$ captures the natural direct effect $E[Y_{a,M_{a^*}} - Y_{a^*,M_{a^*}}]$, that $c_2(a - a^*)$ captures the natural indirect effect $E[Y_{a,M_a} - Y_{a,M_{a^*}}]$, and that their sum measures the total effect $E[Y_{a,M_a} - Y_{a^*,M_{a^*}}] = E[Y_a - Y_{a^*}]$.

Equation 3 is a special case of the more general class of generalized linear MSMs given by

$$g(E[Y_{a,M_{a^*}}]) = c_0 + c_1a + c_2a^* + c_3a \cdot a^*, \quad (4)$$

where g is a link function specifying the requested model for the outcome (e.g., logistic model), and c_3 is an interaction term, which can be included if required. When $c_3 = 0$ and g is the logit link, then $\exp[c_1(a - a^*)]$ captures the natural direct effect odds ratio: $\text{odds}[Y_{a,M_{a^*}} = 1]/\text{odds}[Y_{a^*,M_{a^*}} = 1]$; $\exp[c_2(a - a^*)]$ captures the natural indirect effect odds ratio: $\text{odds}[Y_{a,M_a} = 1]/\text{odds}[Y_{a,M_{a^*}} = 1]$; and their product measures the total effect: $\text{odds}[Y_a = 1]/\text{odds}[Y_{a^*} = 1]$. Further, a value c_3 differing from zero indicates that the magnitude of the direct effect may depend on the natural level at which the mediator is controlled and may thus be the result of an exposure-mediator interaction. Robins and Greenland (6) and Hafeman and Schwartz (9) proposed the terms pure- and total natural (in)direct effects when such interactions are present. Thus, when c_3 is nonzero and g is the logit link, the natural effects of changing exposure from a^* to a are

given by the following:

Pure natural direct effect odds ratio:

$$\exp[c_1(a - a^*) + c_3a^*(a - a^*)];$$

Total natural direct effect odds ratio:

$$\exp[c_1(a - a^*) + c_3a(a - a^*)];$$

Pure natural indirect effect odds ratio:

$$\exp[c_2(a - a^*) + c_3a^*(a - a^*)];$$

Total natural indirect effect odds ratio:

$$\exp[c_2(a - a^*) + c_3a(a - a^*)].$$

When, instead, exposure-covariate interactions are of interest, then the above generalized linear MSMs can be phrased conditional on covariates as

$$\begin{aligned} g(E[Y_{a,M_{a^*}}|C]) &= c_0 + c_1a + c_2a^* + c_3a \cdot a^* + c_4C \\ &\quad + c_5a \cdot C + c_6a^* \cdot C. \end{aligned} \quad (5)$$

Here, c_5 captures the extent to which the direct effect is modified by covariates, and c_6 captures the extent to which the indirect effect is modified by covariates.

The class of generalized linear MSMs encompasses a wide range of models, but the Cox and Aalen models (21), which are important models for survival data, are not included in the class. Cox and Aalen models assume that the hazard function corresponding to the counterfactual survival time, $Y_{a,M_{a^*}}$, can be expressed as

$$\lambda_0(t) \exp(c_0 + c_1a + c_2a^* + c_3a \cdot a^*) \text{ (CoxMSM)} \quad (6)$$

$$\gamma_0(t) + c_1a + c_2a^* + c_3a \cdot a^* \text{ (AalenMSM)}, \quad (7)$$

where $\lambda_0(t)$ and $\gamma_0(t)$ are unspecified baseline hazards. Whenever the outcome is a survival time, we will in addition assume that censoring satisfies the usual assumptions, that is, that censoring is independent of event time (22) conditional on the covariates in the MSM. The rest of the paper is devoted to estimating MSMs that can be written as in equations 4–7.

We propose to estimate the MSMs given by equations 4–7 by the following procedure, which generalizes a proposal by Hong (16) to models for the nested counterfactuals $Y_{a,M_{a^*}}$, corresponding to a wide variety of outcome types. We explain it first for a dichotomous exposure A . Construct a new data set by repeating each observation in the original data set twice and including an additional variable A^* capturing the 2 possible values of the exposure relative to the indirect path. For the first replication of the observation, A^* is set to the actual value of the exposure (that is $A_i^* = A_i$), while for the second replication, A^* is set to the opposite of the actual exposure (i.e., $A_i^* = 1 - A_i$ when A is coded to be 0 or 1). The MSM given by equations 4–7 can now be estimated from the new data set by using standard software by regressing the outcome on the observed exposure A and

the additional variable A^* on the basis of the new data set, weighting each observation in the expanded data set with

$$W_i = \frac{1}{P(A = A_i | C = C_i)} \frac{P(M = M_i | A = A_i^*, C = C_i)}{P(M = M_i | A = A_i, C = C_i)}.$$

The first fraction in these weights ensures that the exposure-outcome association is adjusted for confounding by C . Indeed, the impact of up-weighting observations with a rare combination of exposures and confounders is to create a pseudo-population in which the exposure is no longer associated with C and, thus, there is no residual confounding by C (i.e., mimicking a randomized trial). The second fraction of the weights serves to distinguish between the direct and indirect paths (by correcting for the fact that the observed mediator value may differ from the counterfactual value M_{a^*} that is of interest).

The above approach can easily be implemented in standard software by using the following simple estimation procedure for a dichotomous exposure, which apart from step 3 is completely analogous to estimation of standard MSMs:

1. Estimate a suitable model for the exposure conditional on confounders by using the original data set.
2. Estimate a suitable model for the mediator conditional on exposure and baseline variables by using the original data set.
3. Construct a new data set by repeating each observation in the original data set twice and including an additional variable A^* , which is equal to the original exposure for the first replication and equal to the opposite of the actual exposure for the second replication. In addition, add an identification variable to indicate which data rows originate from the same subject.
4. Compute weights by applying the fitted models from steps 1 and 2 to the new data set. In most software packages, this can be done by using “predict-functionality.”
5. Fit a suitable model to the outcome including only A and A^* (and perhaps their interaction) as covariates and weighted by the weights from the previous step. It can be shown (Web Appendix 3), provided that the exposure and mediator models in steps 1 and 2 are fitted by using a standard maximum likelihood procedure and provided that the mediator model is sufficiently rich so as not to contradict the restrictions imposed by the chosen generalized linear MSM, that conservative confidence intervals can be obtained as the estimate of the natural direct or indirect effect plus/minus 1.96 times a robust standard error, which can be obtained by using software for generalized estimating equations; alternatively, a bootstrap procedure can be used.

It is well established in the literature on MSMs (refer to the article by Robins et al. (18)) that estimators based on inverse probability weights like $1/P(A = A_i | C = C_i)$ can be unstable in samples of small to moderate size, as the weights can become so large that individual observations dominate the estimation. Unless the MSM is saturated,

somewhat better behaving estimators may be obtained by instead using stabilized weights given by

$$W_i^s = \frac{P(A = A_i)}{P(A = A_i | C = C_i)} \frac{P(M = M_i | A = A_i^*, C = C_i)}{P(M = M_i | A = A_i, C = C_i)}.$$

When the MSM includes the covariates C (compare with equation 5), then the following stabilized weights can also be used:

$$W_i^c = \frac{P(M = M_i | A = A_i^*, C = C_i)}{P(M = M_i | A = A_i, C = C_i)}.$$

Note that these do not involve inverse probability weighting by the exposure distribution, because the adjustment for confounding by C now happens via a standard regression adjustment. These weights will thereby typically be much more stable.

The above approach is very flexible because, unlike traditional approaches, it does not work indirectly by combining parameter estimates from standard models for the mediator M and outcome Y . In particular, it can, in principle, be used for any type of outcome, mediator, and exposure and regardless of the choice of MSM and the models for mediator and exposure. However, the approach may lend itself less ideally to the analysis of continuous mediators, because this requires substituting the probabilities $P(M = M_i | A = A_i, C = C_i)$ in the weights by probability densities, which in turn may yield unstable weights (refer to Web Appendix 1 or 2, however, for an application where this was not the case). For categorical exposures A , a minor modification is needed in that one must repeat the original data set as many times as needed to ensure that, for each subject, A^* takes on all the possible values it can take. For continuous exposures, we recommend fitting MSMs conditional on covariates, with corresponding stabilized weights W_i^c to avoid instability due to inverse weighting by the exposure distribution. Here, the user is advised to follow the procedure prescribed for categorical exposures but to replace A^* for subject i by randomly drawn exposures. This can either be done by resampling from the observed exposures or by drawing from a normal distribution with the mean and standard deviation matching the observed exposures. For continuous exposures, a minimum of 5 draws must be made for each original observation.

Web Appendices 3 and 4 contain a mathematical validation of the procedure and of the validity of the robust standard errors, respectively. In addition, Web Appendices 1 and 2 present 2 implementations (one with a binary outcome and one with a survival outcome) of this procedure in both SAS software and R language.

DISCUSSION AND CONCLUSION

It should be noted that the proposed estimators do not exploit all available information in the data and, thus, that more efficient estimators can, in principle, be obtained. Furthermore, their correctness critically hinges on the correctness of the MSMs and the models used for the exposure and

the mediator, except when the MSM is specified conditional on covariates C , in which case correct specification of the exposure model is not required. If only the MSM model is misspecified, then the resulting measures for natural effects can still be interpreted as a “best” approximation; however, this is no longer the case when the exposure or mediator models are misspecified. The current advice must therefore be to conduct a thorough misspecification analysis for the 2 models used in this paper and to evaluate the stability of the weights. Tchetgen Tchetgen and Shpitser (20) and subsequently Zheng and van der Laan (23) proposed estimators that are efficient and multiply robust in the sense that they merely require the correctness of 2 out of 3 models (the 3 models being the model for the exposure, the model for the mediator, and the model for the outcome) to be correct, regardless of which 2 are correct. However, implementation of these estimators is more demanding at present. Work is ongoing to develop alternative estimators that can also be obtained via standard software, are more efficient than the ones proposed in this paper, and also share multiple robustness properties.

The sensitivity of effects separation techniques toward unmeasured confounders is an active area of research (11, 20, 24–26). Web Appendix 5 presents a simulation study assessing the sensitivity of the described approach toward unmeasured confounders in a simple setup of only binary variables. As expected, the assumption of no-unmeasured confounders is found to be critical, the only exception being unmeasured confounding of the exposure-mediator relation, which does not affect estimation of the direct effect.

In summary, this paper has described a simple unified procedure for estimating natural direct and indirect effects. The procedure can be applied to almost any combination of variable types and can be conducted in standard software. Web Appendixes 1 and 2 of the paper provide detailed implementation examples in SAS software and R language.

ACKNOWLEDGMENTS

Author affiliations: Department of Biostatistics, University of Copenhagen, Copenhagen, Denmark (Theis Lange); and Department of Applied Mathematics and Computer Science, Ghent University, Ghent, Belgium (Stijn Vansteelandt, Maarten Bekaert).

T. L. was supported by the Commission of Social Inequality in Cancer (grant SU08004). S. V. was supported by the IAP research network (grant P06/03) from the Belgian government (Belgian Science Policy). M. B. acknowledges support from the Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT vlaanderen).

Conflict of interest: none declared.

REFERENCES

- Baron RM, Kenny DA. The moderator-mediator variable distinction in social psychological research: conceptual,

- strategic, and statistical considerations. *J Pers Soc Psychol.* 1986;51(6):1173–1182.
- Cole SR, Hernán MA. Fallibility in estimating direct effects. *Int J Epidemiol.* 2002;31(1):163–165.
 - Kaufman JS, Maclehose RF, Kaufman S. A further critique of the analytic strategy of adjusting for covariates to identify biologic mediation. *Epidemiol Perspect Innov.* 2004;1(1):4. (doi:10.1186/1742-5573-1-4).
 - VanderWeele TJ, Vansteelandt S. Conceptual issues concerning mediation, interventions and composition. In: *Statistics and Its Interface.* Vol 2. Somerville, MA: International Press; 2009:457–468.
 - Pearl J. *Causality: Models, Reasoning, and Inference.* New York, NY: Cambridge University Press; 2009.
 - Robins JM, Greenland S. Identifiability and exchangeability for direct and indirect effects. *Epidemiology.* 1992;3(2):143–155.
 - Pearl J. Direct and indirect effects. In: *Proceedings of the American Statistical Association Joint Statistical Meetings.* Brentwood, MO: MIRA Digital Publishing; 2005:1572–1581. (Technical report R-273).
 - Petersen ML, Sinisi SE, van der Laan MJ. Estimation of direct causal effects. *Epidemiology.* 2006;17(3):276–284.
 - Hafeman DM, Schwartz S. Opening the Black Box: a motivation for the assessment of mediation. *Int J Epidemiol.* 2009;38(3):838–845.
 - Imai K, Keele L, Yamamoto T. Identification, inference and sensitivity analysis for causal mediation. *Stat Sci.* 2010;25(1):51–71.
 - Imai K, Keele L, Tingley D. A general approach to causal mediation analysis. *Psychol Methods.* 2010;15(4):309–334.
 - VanderWeele TJ. Marginal structural models for the estimation of direct and indirect effects. *Epidemiology.* 2009;20(1):18–26.
 - VanderWeele TJ, Vansteelandt S. Odds ratios for mediation analysis for a dichotomous outcome. *Am J Epidemiol.* 2010;172(12):1339–1348.
 - Lange T, Hansen JV. Direct and indirect effects in a survival context. *Epidemiology.* 2011;22(4):575–581.
 - VanderWeele TJ. Causal mediation analysis with survival data. *Epidemiology.* 2011;22(4):582–585.
 - Hong G. Ratio of mediator probability weighting for estimating natural direct and indirect effects. In: *Proceedings of the American Statistical Association, Biometrics Section.* Alexandria, VA: American Statistical Association; 2010: 2401–2415.
 - Pearl J. *The Mediation Formula: A Guide to the Assessment of Causal Pathways in Non-Linear Models.* Los Angeles, CA: University of California; 2011. (Technical report R-363).
 - Robins JM, Hernán MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology.* 2000;11(5):550–560.
 - van der Laan MJ, Petersen ML. Direct effect models. *Int J Biostat.* 2008;4(1):1–27.
 - Tchetgen Tchetgen EJ, Shpitser I. Semiparametric estimation of models for natural direct and indirect effects. Berkeley, CA: bepress; 2011. (Harvard University Biostatistics Working Paper Series. Working Paper 129).
 - Aalen O. A model for non-parametric regression analysis of counting processes. In: Klonicki W, Kozek A, Rosinski J, et al. *Lecture Notes in Statistics-2: Mathematical Statistics and Probability Theory.* New York, NY: Springer-Verlag; 1980:1–25.

22. Martinussen T, Scheike TH. *Dynamic Regression Models for Survival Data*. New York, NY: Springer; 2006.
23. Zheng W, van der Laan MJ. Targeted maximum likelihood estimation of natural direct effect. Berkeley, CA: bepress; 2011. (UC Berkeley Division of Biostatistics Working Paper Series. Working Paper 288).
24. Van der Weele TJ. Bias formulas for sensitivity analysis for direct and indirect effects. *Epidemiology*. 2010;21(4):540–551.
25. Hafeman DM. Confounding of indirect effects: a sensitivity analysis exploring the range of bias due to a cause common to both the mediator and the outcome. *Am J Epidemiol*. 2011;174(6):710–717.
26. Christensen KB, Labriola M, Lund T. Explaining the social gradient in long-term sickness absence: a prospective study of Danish employees. *J Epidemiol Community Health*. 2008; 62(2):181–183.