Statistical issues in life course epidemiology

Bianca L. De Stavola[1], Dorothea Nitsch[1], Isabel dos Santos Silva[1], Valerie McCormack[1], Rebecca Hardy[2], Vera Mann[1], Tim J. Cole[3], Susan Morton[1,4], David A Leon[1]

[1] Department of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine

[2] MRC National Survey of Health and Development, Department of Epidemiology, Royal Free and University College Medical School

[3] Centre for Paediatric Epidemiology and Biostatistics, Institute of Child Health, University College London

[4] (current address) Liggins Institute and School of Population Health, University of Auckland, New Zealand

Short running title: Life course issues

Words count : abstract = 200; text = 3,986

Abbreviations : CI = confidence interval; EM = expectation-maximization; ML: maximum likelihood; MRC: Medical Research Council; NSHD: National Survey of Health and Development; OR = odds ratio; SD = standard deviation; SEM: structural equation model.

ABSTRACT

There is growing recognition that the risk of many diseases in later life, such as type-2 diabetes or breast cancer, is affected by adult as well as early life variables, including those operating prior to conception and during the pre-natal period. Most of these risk factors are correlated because of common biological and/or social pathways, while some are intrinsically ordered over time. The study of how they jointly influence later (`distal') disease outcomes is referred to as life course epidemiology. This area of research raises several issues that are relevant to the current debate on causal inference in epidemiology. The authors give a brief overview of the main analytical and practical problems and consider a range of modelling approaches, their differences being determined by the degree with which relationships present (or presumed) among the correlated explanatory variables are explicitly acknowledged. Standard multiple regression (i.e. univariate) models are compared to multivariate models where several outcomes are jointly specified. Issues arising from measurement error and missing data are addressed. Examples originated from two UK cohorts are used to illustrate alternative modelling strategies. It is concluded that more than one analytical approach should be adopted to gain more insight into the underlying mechanisms.

KEYWORDS: Correlated data, intermediate exposures, life course, path analysis, structural equation models

In the last decade there has been a growing realization that prenatal and early life biological and social factors may play an important role in the aetiology of many later-life conditions (1-9), in addition to - or in synergism with - those of well-established adult exposures. This field of enquiry is now referred to as life course epidemiology (10).

Studying complex inter-relationships of biological and social variables over time requires longitudinal information spanning broad periods of life. It also raises analytic problems because temporal and -possibly- causal hierarchies among the exposures need to be taken into account (11,12). For example, breast cancer risk factors operate at a number of stages in the life course (Figure 1). They also may influence each other, e.g. childhood weight is inversely correlated with on timing of puberty (13) and -directly or indirectly- on later obesity (14). Age at puberty and adult obesity affect breast cancer risk (15) and thus may mediate the effect of childhood weight. If standard multiple regression models are used to study these variables, the effect of childhood weight would be estimated holding all other exposures constant, thus missing the life course perspective, as we show below. Alternatively, if a multivariate approach is used, where for example age at menarche, adult obesity and breast cancer risk are joint outcomes, their interrelationships would be explicitly estimated, albeit within an assumed multivariate structure.

We discuss these problems by introducing a sequence of increasingly complex models to deal with the temporal and causal hierarchies among the exposure variables. We compare them in terms of interpretability as well as flexibility in dealing with missing data and measurement errors, both of which are common features in life course studies. We illustrate issues and models with two examples: the first studies intergenerational influences on size at birth using data from a Scottish cohort of children recruited in 1962; the second examines childhood height and its impact on adult leg length with data from a UK birth cohort of women born in 1946. Our aim is to present a broad analytical framework for

studies in life course epidemiology and to highlight its relevance to the current debate on causal modelling (16-18).

MODELS FOR DISTAL OUTCOMES

When a structure among the exposure variables is known or presumed, a distinction can be made between variables that act at the inception (`background') or in the middle (`intermediate') of the process which leads to the main (`distal') outcome of interest (16). In the breast cancer example described above childhood weight is a background variable, age at menarche and adult obesity intermediate variables and breast cancer the distal outcome.

Statistical models can only offer simplified representations of reality (19). We classify those relevant in life-course epidemiology according to the degree to which they explicitly acknowledge the relations among their components.

Univariate models

Typically univariate models express the expectation of the outcome of interest, E(Y) (or a suitable transformation $g(.)$), as a function of several explanatory variables, $X_1, X_2, \ldots, X_k$,

$$g(E(Y)) = ß_0 + ß_1 X_1 + ß_2 X_2 + \ldots + ß_k X_k \qquad (1)$$

where $g(.)$ is the link function used in generalised linear models (20) and the coefficient $ß_j$ $(j=1,\ldots,k)$ represents the effect attributed to one unit increase in X_j when all the other variables are held constant. If $g(.)$ is the identity function, and Y is continuous, equation (1) represents a linear regression model.

A univariate model can be fitted with only background variables, thus avoiding the inclusion of exposures that may be on the causal pathway. This was the approach adopted in the early studies

linking birth weight with coronary heart disease (21). Its estimated coefficients thus measure the effect of each background variable, controlled for that of the others. The longer the time-gap between background variables and distal outcome, however, the greater is the possibility of intervening modifying effects (see the debate surrounding the foetal programming hypothesis: 1-3,22-24).

If all types of exposures (background and intermediate) are included in the same univariate model, the resulting regression coefficients would measure mutually adjusted effects, i.e. effects of background variables not mediated via the intermediate variables and effects of intermediate variables conditional on the background ones. More specifically if, for example, the background variable X_1 influences the intermediate variable X_k, $ß_1$ would only capture the effect of X_1 on Y that is not mediated via X_k.

A special setting which involves the joint analysis of background and intermediate variables arises when repeated measures of the same exposure are taken over time. In this case, the first available measure acts as background for all the following ones. Consider the influence of childhood anthropometric variables on adult obesity (14). Results obtained after including all childhood measurements in the same univariate regression model for the distal outcome (obesity) are difficult to interpret, especially if the measures are taken close to each other in time. With two repeated body size measures, say Z_1 and Z_2, taken at times $t_1<t_2$, the model

$$g(E(Y)) = ß_0 + ß_1 Z_1 + ß_2 Z_2 \qquad (2a)$$

could be re-written as either

$$g(E(Y)) = ß_0 + (ß_1 + ß_2) Z_1 + ß_2 (Z_2 - Z_1) \qquad (2b)$$

or

$$g(E(Y)) = \beta_0 + (\beta_1 + \beta_2)\, Z_2 - \beta_1\, (Z_2 - Z_1) \qquad (2c)$$

where the difference (Z_2-Z_1) represents the change in the explanatory variable, e.g. body size, occurring from time t_1 to time t_2.

The three equations (2a)-(2c) are different parameterizations of the same model (22,23). When Z_2 is replaced by (Z_2-Z_1), the conditional effect of Z_1 on the transformed dependent variable, g(E(Y)), changes from β_1 to $(\beta_1+\beta_2)$ (equation (2b)). A similar switch is observed for the effect of Z_2 when it is conditioned on (Z_2-Z_1) (equation (2c)). The conditional effect of the difference (Z_2-Z_1) is either β_2 or $-\beta_1$, depending on whether you condition on the first or second measure, respectively. So different interpretations are possible depending upon the conditioning variable, as originally discussed by Cole and colleagues (22,25,26) and revisited by others (23,24).

A model with several repeated measures, $Z_1, Z_2, .., Z_K$, taken at times $t_1 < t_2 < ... < t_K$, can be re-parameterized in terms of the first one, Z_1, and all subsequent consecutive increments, $(Z_j - Z_{j-1})$, $j=1,...,K$. Thus equation (2b) becomes:

$$g(E(Y)) = b_0 + \left(\sum_{j=1}^{K} b_j\right) Z_1 + \left(\sum_{j=2}^{K} b_j\right)(Z_2 - Z_1) + \left(\sum_{j=3}^{K} b_j\right)(Z_3 - Z_2) + ... + b_K (Z_K - Z_{K-1}) \qquad (2d)$$

and similarly for equation (2c). In (2d) the coefficient for Z_1 is the sum of all conditional effects associated with $Z_1, Z_2, Z_3, ..., Z_K$, i.e. is the cumulative effect of increasing each of the Z_j by one unit. Such increases would happen, for example, when a unit change at time t_1 has irreversible effects on all following values of Z. For example one extra cm in height at age t_1 has an impact on all future height values of a child, since his/her trajectory is shifted upwards from then on, holding everything else constant, and therefore a cumulative impact on the outcome. Similarly the coefficient for each increment $(Z_j-Z_{j-1}), (j=2,\ldots,K)$, captures the effect of increasing Z during the j^{th} interval, with that change

shifting all subsequent *Z* values. An alternative parameterization, often used in life course epidemiology (13,14), replaces the increments with the equivalent changes per unit time (`velocities'), i.e. $(Z_j - Z_{j-1})/(t_j - t_{j-1})$. In this case the coefficients for the j^{th} period are $(t_j - t_{j-1})$ times the coefficients for $(Z_j - Z_{j-1})$.

A graphical approach – the "life course plot"- may help interpreting the impact of each repeated measure (26). It involves plotting the conditional regression coefficients against the times when measures were taken (after standardisation to make these coefficients comparable). When the coefficients switch sign at some time t_j, as in Figure 2, there is evidence that changes during (t_{j-1}, t_j) have an impact on the outcome of interest.

Multivariate models

Multivariate models deal with several outcomes simultaneously. In this context they would explicitly define a presumed process underlying intermediate and distal outcomes. For example, variables may be arranged as in Figure 3a, where Y is the distal outcome and X_1, X_2 and X_3 the explanatory variables. In this diagram, X_3 is assumed to be directly affected by X_1 and X_2 and thus is an intermediate variable, while X_1 and X_2 are background variables. Its algebraic equivalent is a system of simultaneous equations, with as many equations as there are intermediate and distal outcomes, which takes the name of path analysis (27):

$$\begin{cases} g_y(E(Y)) = ß_0 + ß_1 X_1 + ß_2 X_2 + ß_3 X_3 \\ g_3(E(X_3)) = a_0 + a_1 X_1 + a_2 X_2 \end{cases} \qquad (3)$$

Here, as before, $E(.)$ stands for expectation while $g_y(.)$ and $g_3(.)$ are link functions. When $g_y(Y)$ and $g_3(X_3)$ are conditionally normally distributed estimation can be carried out by maximum likelihood (ML) or, avoiding the normality assumptions, by 3-stage least squares (28). The indirect effects of a background variable, X_1 say, on the distal outcome can be computed by multiplying the

standardised coefficients (i.e. the effects for 1 standard deviation (SD) increase in X_i) found along each of the paths leading from X_i to Y and then summing them (18, 29). In Figure 3a, there is only one indirect path from X_1 to Y, via X_3. Thus, if X_1 and X_3 are standardized and the path model is correct (30), the indirect effect of X_1 is a_1 multiplied by $ß_3$.

When Y and X_3 are not normally distributed, penalized maximum likelihood (31,32) or non-parametric maximum likelihood estimation are used, with the latter recently suggested to improve the estimation properties (33).

When some of the variables are proxies for factors that could not be (or were not) measured precisely, latent variables can be introduced within this framework. For example, several dietary variables may be available via a Food Frequency Questionnaire aiming to measure `usual diet' (34). Thus each of them is a proxy, or `manifestation', of an unmeasurable construct that nevertheless is of interest. Similarly repeated height observations during childhood are manifestations of an underlying growth pattern.

In Figure 3b three variables, X_1, X_2 and X_3, act as proxy (or `manifest') measures for the unmeasurable/ unmeasured variable U (the convention is to use squares for observed variables and circles for latent variables). The effect of U on Y can be estimated by specifying how the three proxies are linearly related to U and also how U is related to Y:

$$\begin{cases} E(X_1) = \mu_1 + \ ?_1 U \\ E(X_2) = \mu_2 + \ ?_2 U \\ E(X_3) = \mu_3 + \ ?_3 U \\ g_Y(E(Y)) = ß_0 + \ ß_1 U \end{cases} \qquad (4)$$

Usually the latent variable U is assumed to be normally distributed and the parameters μ_1, μ_2 and μ_3 are set to be zero. The link function for Y, $g_y(E(.))$, however could be of any form. Other observed variables could be influencing Y; in this case the last equation in (4) would have an additional term $ß_4 X_4$.

The first three equations in (4) define the *measurement* part of the model because U is not observed, but proxied by X_1, X_2 and X_3. The fourth equation defines the *structural* part, i.e. the relation between the unmeasured variable U and the distal outcome Y (35). Together, the measurement and structural models form a structural equation model (SEM; 29).

Because U is not directly observed and does not usually have a quantifiable metric, its influence on the manifest variables can only be measured in terms of an arbitrary metric. One convention is to use the first of the proxy variables as reference and thus adopt its metric, e.g. that of X_1, so that the effect of the latent variable on Y becomes expressed in terms of X_1 units. Alternatively, the variance of the latent construct is fixed to be 1 and its effect on Y estimated in terms of 1 SD change in the latent variable.

As for path analysis, estimation can be carried out by maximum, penalized maximum (31,32) or non-parametric maximum likelihood (33), depending on the link functions. Generalizations to include more than one latent variable (and relationships among them) are straightforward, although issues of identifiability constrain their numbers (29). Generalizations to discrete latent variables are also within the scope of these models. They involve the concept of latent classes, the probability of belonging to each of them being determined by a higher level latent continuous factor (36, 37), with estimation carried out by maximum likelihood with the expectation-maximization (EM) algorithm (36, 38,39). Multivariate models such as these can be fitted in M-Plus (40) and Stata (33).

Data quality issues

A major difficulty that arises when analysing life course studies derives from varying data quality. Because the focus is on different time periods, data from multiple sources (including routine data, e.g. cancer registries) are merged although their variable definition may vary, as well as their completeness. The number of subjects with complete data on any subset of the variables of interest can therefore be reduced to a small fraction of the total while the precision of a variable may depend on when it was collected, for example because of changes in units (as for birth weight, measured in pounds or kg). Thus measurement errors and missing values affect life course studies to a greater extent than standard observational studies.

Several methods are available to deal with measurement errors (41). For example, when data on two or more proxy variables for an exposure of interest are collected, calibration methods can be used within a univariate approach (42,43). Alternatively, as described above, multivariate models that include latent variables can be fitted (44). When the data are affected by missing values, analyses based on complete-records (via univariate or multivariate models) are rarely appropriate because the incorrect assumption of a "missing completely at random" mechanism would lead to biased results (45). If missingness can be assumed to be "at random" (i.e. MAR) either of two closely related (46) approaches can be used: imputation methods with univariate models (47), and ML plus the EM algorithm (36), or Bayesian simulations (48), with multivariate models. When instead data are suspected to be systematically missing because of unmeasured factors, extensive sensitivity analyses- either in univariate or multivariate models- should be performed (49-52).

EXAMPLES AND DATA

Two examples arising in the analysis of two UK cohorts will be used for illustration. The first investigates how maternal and grand-maternal factors influence the size at birth of an offspring using the *Children of the 1950s Study*. This cohort includes all people who in 1962 participated in a reading survey while attending primary school in Aberdeen (Scotland; 53,54). Data were collected from the participants' obstetric records and on characteristics of their parents and grand-parents. In 1999-2000, 4497 female study members (78% of the total) were anonymously linked to Scottish maternity records leading to birth data on their offspring. Thus information on three generations (the grandparents, denoted G0, the study participants, G1, and their offspring, G2) can be studied.

The second example focuses on how adult leg-length, which has been used in cancer and cardiovascular epidemiology as a marker of childhood environmental factors (55,56), is determined by different periods of childhood growth. The leg-length of participants in the *Medical Research Council National Survey of Health and Development* (*NSHD*) was measured by a trained nurse when aged 43 years. The *NSHD* is a socially stratified birth cohort which includes 2547 women and 2815 men born during the week 3-9 March 1946 (57-59) and followed prospectively, with childhood height measured at ages 2, 4, 6, 7, 11, and 14/15 years by trained personnel.

RESULTS

Example 1: Intergenerational influences on size at birth

We aimed to investigate how strongly intergenerational factors influence a baby's size at birth (defined as birth weight standardised for gestational age) using data from the *Children of the 1950s* Study (Figure 4;54). For illustration we consider only biological G1 and G0 factors (birth size of the

mother, and height and parity of mother and grand-mother) and restrict the analyses to 1724 first singleton baby girls.

Fitting a univariate linear regression model for G2 birth size on all potential explanatory variables shows that all maternal factors have positive effects, unlike the grand-maternal ones (Table 1a). Thus offspring (G2) size at birth is larger when the mother (G1) is taller, holding constant all other variables, but is reduced (although not significantly) when, again holding all other variables constant, her grand-mother (G0) is taller or had more children. The negative G0 height effect should be interpreted in conjunction with the positive G1 effect, since these two variables are *de facto* repeated measures of the same variable. A clearer interpretation would focus on their difference, and conclude that the taller a G1 woman is relative to her mother the larger her offspring (estimated $ß_2$=0.18 in equation (2b)).

By instead fitting the path analysis model equivalent to Figure 4 we can deal with these relationships simultaneously (Table 1b). The estimated direct effect coefficients corresponding to the arrows leading to the multivariate outcomes (within double boxes) show that G1 adult height increases with G0 adult height and G1 birth size but decreases with increasing G0 parity (i.e. was lower in larger families) while G1 birth size increases with increasing G0 parity and G0 adult height. By multiplying the standardized parameters along the relevant pathways, the indirect effects on G2 birth size of G0 parity, G0 adult height and G1 birth size can be estimated (Table 2). They show that, although G0 height has a negative direct effect (i.e. not mediated), its indirect effect via G1 birth size and G1 height is positive, leading to a positive and significant total effect (0.13= 0.49*0.18 (via G1 height) +0.20*0.19 (via G1 birth size)+0.20*0.19*0.18 (via G1 birth size and height); Table 2). So, although the univariate model gave an insight into intergenerational direct effects, the path model led to a more comprehensive summary of their inter-relationships.

Example 2: Childhood growth and adult leg-length

We used data on 2349 the female participants in *NSHD,* for whom at least one childhood height or adult leg-length was available, to investigate which childhood periods are most associated with adult leg-length. A series of univariate regression models for adult leg-length were fitted adding each childhood height measure one at a time, starting from age 2 years. Because of missing values these models are based on different numbers of observations (Table 3a).

At first sight, the results are difficult to interpret because of the changing size and, occasionally, sign of the parameters obtained for the same height measure in different models. The only systematic feature of these models is the consistently larger size and significance of the estimated coefficient corresponding to the oldest age. Model 4 is however the exception. Here the oldest age, 11 years, reflects stage of sexual maturation as much as linear growth- and stage of maturation is a poor predictor of adult leg-length. Thus it is not surprising that height at this age has a weak effect on adult leg-length when conditioned on earlier stature.

In the model including all available childhood measures (model 5) the effects of height at age 2 and 11 are negative, showing that, conditionally on all other childhood measures, being shorter at these ages leads to longer adult leg-length. This can also be seen when plotting the equivalent standardised coefficients, as suggested by Cole (26; Figure 5). When heights are replaced by height increments (Table 3b), the coefficients in each newly fitted model are the sum of the coefficients in the original one (as shown in equation (2d)). For example in model 5 the coefficient for height increments between age 6 and 7 years is precisely 0.493=0.148-0.102+0.447, i.e. is the sum of the conditional height effects at ages 7, 11 and 15. Thus the coefficients in Table 3b capture the cumulative effect of a shift in height at one age as it impacts on a girl's height at all subsequent ages.

Given the similarity of the coefficients for the earliest height differences, the model can be simplified to include only the intervals 2 to 7 years, 7 to 11 years, and 11 to 15 years (Tables 4a-4b). To obtain comparable measures of effects the model with yearly height velocities is also reported (Table 4c). Here the estimated coefficients are multiples of those found when using height increments: for example the coefficient for height velocity between age 2 and 7 years is five times that for the equivalent height difference and 0.55=0.19-0.10+0.45 is the sum of the conditional height effects.

An equivalent multivariate analysis would assume that a girl's growth profile is determined by a latent process that influences her adult leg-length. Thus we parameterized the growth process in terms of average height at age 2, and height velocities between the ages of 2 and 7, 7 and 11, and 11 to 15 years (Figure 6). These are latent variables which are manifested by the observed heights at 2, 4, 6, 7, 11 and 15 years (the `measurement model'), where the latent variables are equivalent to random coefficients in a generalized linear mixed model (60-62). The structural part of this multivariate model defines instead how the latent variables influence adult leg-length (Figure 7).

The measurement and structural models were jointly fitted, the first giving estimated mean growth parameters (Table 5) which were consistent with both observed values (Table 4c) and standard growth charts (63,64). The structural model gave estimates of the effects of the growth parameters on adult leg-length, which were similar to those found by the univariate model with height at age 2, 7, 11 and 15 years. The mainly small numerical differences between univariate and multivariate results are to be attributed to two main factors. Firstly the multivariate model was fitted using all height measures, with the consequent impact on average height at age 2. Secondly, the multivariate model dealt with the measurement error that arises when height velocities are calculated from observed data by specifying them as latent variables. Thus the confidence intervals for the multivariate parameters are wider, better reflecting the uncertainty about the underlying growth process.

For comparison the multivariate estimates were obtained on the same subset of women that contributed to the univariate model (N=794). However, assuming that the missing height and leg-length values occurred at random (45), the multivariate model can be re-fitted using the 2349 women who had at least one height or leg measure, via the EM algorithm available in M-Plus (Table 5b; 40). The results are marginally different from those found earlier, reflecting the fact that the 794 women in the initial analyses were on average taller at younger ages, grew slower from 11 to 15 years, and had shorter leg-length than the rest (Table 6).

DISCUSSION

In this paper we have described the issues arising when explanatory variables are closely associated because of underlying temporal or biological processes, a frequent feature of life course studies. Univariate and multivariate models have been compared using examples drawn from our work in cardiovascular and cancer epidemiology. These were chosen for their relative simplicity, the purpose of the analyses being illustrative, so that thorough epidemiological investigations were restrained and technical details avoided. Other, more complex, applications can be found in the life course literature (e.g.,14,65,66).

We have used the classification of univariate and multivariate models to contrast two main analytical approaches. Univariate models are relatively easy to apply but also to misinterpret when the conditioning variables are ignored. In contrast, multivariate models may seem ideally suited to deal with life course problems because they explicitly specify the presumed causal and temporal mechanisms for the distal outcome. Further missing data problems can be dealt with directly, if a MAR assumption is appropriate, and measurement error problems by specifying latent variables within a SEM. However

several alternative model specifications (`structures') might be appropriate for a particular application. Thus the choices may be too subjective, especially because formal comparisons are problematic (67).

Whatever approach is adopted the main issue is how to deal with life course dependencies while at the same time considering the impact of unmeasured - or poorly measured - factors that influence the pathway of interest. This is not a new topic in epidemiology (68-70), although it has recently become the focus of renewed interest, as demonstrated by the current debate on causal modelling (18,24,71-78) and on the role of statistics in causal inference (79-82). Inspired by that debate we have used the distinction between univariate and multivariate modelling in the context of life course studies, mirroring the distinction between descriptive and causal modelling (68). Indeed SEMs could be viewed as algebraic representations of causal beliefs (18). It must be stressed however that either approach is prone to misspecifications and thus should not be singly relied upon for causal inference (16). To achieve robust conclusions more than one analytical approach should be adopted with the results compared and inconsistencies investigated, thus carrying out sensitivity analyses in the broader sense (83).

ACKNOWLDGEMENTS

We wish to thank Professors Michael Wadsworth and Diana Kuh for access to the MRC *NSHD* data and Michael Hills and Jonathan Sterne, for invaluable comments on an earlier draft.

GRANTS

This work was conducted within the framework funded by the Medical Research Council Co-operative Group grant on "Life-course and trans-generational influences on disease risk" (Grant no. G9819083).

References

1. Barker DJP. Mothers, Babies and Health in Later Life. 1998; Edinburgh: Churchill Linvingstone.
2. Leon DA, Lithell HO, Vågerö D, Koupilová I, Mohsen R, Berglund L, Lithell U-B, and McKeigue PM. Reduced fetal growth rate and increased risk of ischaemic heart disease mortality in 15 thousand Swedish men and women born 1915-29. BMJ 1998; 317: 241-245.
3. McCormack VA, dos Santos Silva I, De Stavola BL, Mohsen R, Leon DA, and Lithell HO. Fetal growth and subsequent risk of breast cancer: results from long term follow up of Swedish cohort. BMJ 2003; 326 : 248-253.
4. Lithell HO, McKeigue PM, Berlung L and 3 others. Relationship of size at birth to non-insulin-dependent diabetes and insulin levels in men aged 50 to 60 years. BMJ 1996; 312: 406-410.
5. Moore SE, Cole TJ, Collinson AC, Poskitt EM, McGregor IA, Prentice AM. Prenatal or early postnatal events predict infectious deaths in young adulthood in rural Africa. Int J Epidemiol 1999; 28: 1088-1095.
6. Hall AJ, Yee LJ, Thomas SL. Life course epidemiology and infectious diseases. Int J Epidemiol 2002; 31: 300-301.
7. Krieger N. Themes in social epidemiology in the 21th century: an ecosocial perspective. Int J Epidemiol 2001; 30: 668-677.
8. Nagin DS, Tremblay RE. Parental and early childhood predictors of persistent physical aggression in boys from kindergarten to high school. Arch Gen Psychiatry 2001; 58: 389-394.
9. Beebe-Dimmer J, Lynch JW, Turrell G and 3 others. Childhood and adult socioeconomic conditions and 31-year mortality risk in women. Am J Epidemiol 2004; 159: 481-490.

10. Ben-Shlomo Y, Kuh D. [editorial] A life course approach to chronic disease epidemiology: conceptual models, empirical challenges and interdisciplinary perspectives. Int J Epidemiol; 31: 285-293.
11. Hallqvist J, Lynch J, Bartley M,et al. Can we disentangle life course processes of accumulation, critical period and social mobility? An analysis of disadvantaged socio-economic positions and myocardial infarction in the Stockholm Heart Epidemiology Program. Soc Sci Med. 2004; 58: 1555-1562.
12. Goldberg GR, Prentice AM. Maternal and fetal determinants of adult diseases. Nutr Rev. 1994; 52: 191-200.
13. dos Santos Silva I, De Stavola BL, Mann V, Kuh D, Hardy R, Wadsworth M. Prenatal factors, childhood growth trajectories and age at menarche. Int J Epidemiol 2002; 31: 405-412.
14. Naumova EN, Must A, Laird NM. Tutorial in biostatistics: Evaluating the impact of `critical periods' in longitudinal studies of growth using piecewise mixed effects models. Int J Epidemiol 2001; 30: 1332-1341.
15. Kuller LH. The etiology of breast cancer- from epidemiology to prevention. Public Health Review 1995; 23: 157-213.
16. Cox DR, Wermuth N. Causality: a statistical view. J Int Stat Inst 2004; 72: 285-305.
17. Pearl, J. Direct and indirect effects. Proceedings of the Seventh Conference on Uncertainty in Artificial Intelligence. 2001. San Francisco, CA: Morgan Kaufmann, 411-420.
18. Greenland S, Brumback B. An overview of relations among causal modelling methods. Int J Epidemiol 2002; 31: 1030-1037.
19. Cox DR. Role of models in statistical analysis. Statistical Science 1990; 5: 169-174.
20. McCullagh P, Nelder J. Generalized Linear Models. 1980; London: Chapman and Hall.

21. Barker DJP, Osmond C, Winter PD, Margetts B, Simmonds SJ. Weight in infancy and death from ischaemic heart disease. Lancet 1989; 2: 577-580.
22. Lucas A, Fewtrell MS, Cole TJ. Fetal origins of adult disease- the hypothesis revisited. BMJ 1999; 319: 245-249.
23. Tu Y-K, West R, Ellison GTH, Gilthorpe MS. Why evidence for the fetal origins of adult disease might be a statistical artefact: the "reversal paradox" for the relation between birth weight and blood pressure in later life. American Journal of Epidemiology 2005; 161: 27-32.
24. Weinberg CR. Invited commentary: Barker meets Simpson. American Journal of Epidemiology 2005; 161: 33-35.
25. Horta BL, Barros FC, Victora CG, Cole TJ. Early and late growth and blood pressure in adolescence. J Epidemiol Community Health 2003; 57: 226-230.
26. Cole TJ. Modeling postnatal exposures and their interactions with birth size. J Nutr 2004; 134: 201-204.
27. Wright S. On the method of path coefficients. Annals of Mathematical Statistics. 1934; 5: 161-215.
28. Zellner A, Theil H. Three stage least squares: simultaneous estimate of simultaneous equations. Econometrica 1962; 29: 63-68.
29. Bollen KA. Structural Equations with Latent Variables. 1989; New York, NY: Wiley.
30. Cole SR, Herman MA. Fallibility in estimating direct effects. Int J Epidemiol 2002; 31: 163-165.
31. Breslow NE, Clayton DG. Approximate inference in generalised linear mixed models. JASA 1993; 88: 9-25.
32. Muthén BO. A general structural equation model with dichotomous, ordered categorical and continuous latent indicators. Psychometrika 1984; 49: 115-132.

33. Rabe-Hesketh S, Skrondal A, Pickles A. Reliable estimation of generalised linear mixed models using adaptive quadrature. Stata Journal 2002; 2: 1-21.
34. Willett WC. Nutritional epidemiology. 2nd ed. 1998; New York, NY: Oxford University Press.
35. Clayton DG. Models for the analysis of cohort and case-control studies with inaccurately measured exposures. In Statistical Models for Longitudinal Studies on Health. Dwyer HD, Feileib M, Lippert P et al (eds). 1992. Oxford: Oxford University Press.
36. Muthen B, Shedden K. Finite mixture modeling with mixture outcomes using the EM algorithm. Biometrics 1999; 55: 463-469.
37. Bollen KA. Structural equation models. In Encyclopaedia of Biostatistics. Armitage P, Colton T (eds), 1998; J. Wiley: Chichetser, p 4263-4272.
38. Dempster AP, Laird NM, Rubin DB. Maximum likelihood for incomplete data via the EM algorithm (with Discussion). JRSS, B. 1977; 39:1-38.
39. Muthén BO. Latent variables mixture modeling. In Marcoulides, Schumacker RE (eds.) New Developments ad Techniques in Structural Equation Modeling. 2001; Lawrence Erlbaum Ass: 1-33.
40. Muthén LK, Muthén BO. Mplus. Statistical Analysis with Latent Variables. User's Guide. 1998-2004, Los Angeles CA: Muthén & Muthén.
41. Carroll RJ. Measurement error in epidemiological studies. In Encyclopedia of Epidemiologic Methods. Gail MH, Benichou J. (eds) 2000. J Wiley: Chichester.
42. Carroll RJ, Stefanski LA. Approximate quasi-likelihood estimation in models with surrogate predictors. JASA. 1990; 85:652-663.
43. White I, Frost C, Tokunaga S. Correcting for measurement error and continuous variables using replicates. Statistics in medicine 2001; 20: 3441-3457.

44. Rabe-Hesketh S, Skrondal A, Pickles A. Maximum likelihood estimation of generalised linear models with covariate measurement error. The Stata Journal. 2003; 3:385-410.
45. Little RJ, Rubin DB. Statistical Analysis with Missing Data. 1987. Wiley: New York.
46. Little RJ. Missing Data. In Encyclopaedia of Epidemiological Methods. Gail MH, Benichou J (eds). 2000. John Wiley: New York.
47. Shafer JL. Analysis of Incomplete Multivariate Data. 1997.Chapman and Hall: London.
48. Gilks WR, Richardson S, Spiegelhalter DJ (eds). Markov Chain Monte Carlo in Practice. 1996. London: Chapman and Hall.
49. Molenberghs, G., Kenward, M.G., Goetghebeur, E. (2001). Sensitivity analyses for incomplete contingency tables: the Slovenian plebiscite case. Appl. Statist., **50**, 15-29.
50. Kenward, M.G. (1998). Selection models for repeated measurements with non-random drop-out: an illustration of sensitivity. Statist. Med., **17**, 2723-2732.
51. Manski, C. (1989) Anatomy of the selection problem. Journal of Human Resources, **24**, 343-360.
52. Verzilli, C.J., Carpenter, J.R. (2002) Assessing uncertainty about parameter estimates with incomplete repeated ordinal data. Statistical Modelling, **2**, 203-215.
53. Batty GD , Morton SMB, Campbell D, et al. The Aberdeen Children of the 1950s cohort study: background, methods, and follow-up information on a new resource for the study of life-course and intergenerational effects on health. Paedriatr Perinat Epidemiol 2004; 18: 221-239.
54. Morton S. Intergenerational determinants of size at birth. PhD Thesis. 2002; London:University of London.
55. Gunnell D, Okasha M, Holly J, et al. Height and cancer risk: a systematic review of prospective studies and possible mechanisms. Epidemiol Rev 2001; 23: 296-325.

56. Langenberg C, Hardy R, Kuh D, et al. Influence of height, leg and trunk length on pulse pressure, systolic and diastolic blood pressure. Journal of Hypertension 2003; 21:537-543.
57. Wadsworth ME, Mann SL, Rodgers B, Kuh DJ, Hilder WS, Yusuf EJ. Loss and representativeness in a 43 year follow up of a national birth cohort. J Epidemiol Community Health 1992; 46: 300-304.
58. Wadsworth MEJ, Butterworth SL, Hardy RJ, et al. The life course prospective design: an example of benefits and problems associated with study longevity. Social Science Med 2003; 57: 2193-2205.
59. Wadsworth MEJ, Butterworth SL, Montgomery SM, et al. In: Ferri E, Bynner J, Wadsworth MEJ, eds. Changing Britain, Changing Lives. 2003; London: Institute of Education Press; p 207-236.
60. Laird NM and Ware JH. Random-effects models for longitudinal data. Biometrics 1982; 38: 963-974.
61. Goldstein H. Multilevel Statistical Models. 1995; Edward Arnold: London.
62. Muthén BO, Khoo ST. Longitudinal studies of achievement growth using latent variable modeling. Learning and Individual Differences 1998; 10: 73-101.
63. Tanner JM. Foetus into Man. Physical Growth from Conception to Maturity. 2nd edition. 1989; Ware, England: Castlemead Publications.
64. National Center for Health Statistics. 2000 CDC growth charts: US. www.cdc.gov/growthcharts; September 2004.
65. De Stavola BL, dos Santos Silva I, McCormack V, Hardy RJ, Kuh DJ, Wadsworth MEJ. Childhood growth and breast cancer. Am J Epidemiol 2004; 159:671-682.

66. Nitsch D, De Stavola BL, Morton S, Leon D. Linkage bias in estimating the association between childhood exposure and propensity to become a mother: an example of simple sensitivity analyses. (JRSS, A; under revision)
67. Jöreskog KG. Testing structural equation models. In Testing Structural Equation Models. Bollen KA, Scott Long J (eds). 1993; Newbury Park, London: Sage Publications.
68. Robins JM, Greenland S. The role of model selection in causal inference from nonexperimental data. Am J Epidemiol 1986; 123: 392-402.
69. Phillips AN, Davey Smith G. How independent are "independent" effects? Reative risk estimation when correlated exposures are measured imprecisely. J Clin Epidemiol 1991; 11: 1223-1231.
70. Victora CG, Huttly SR, Fuchs SC, Olinto MT. The role of conceptual frameworks in epidemiological analysis: a hierarchical approach. Int J Epidemiol 1997; 26: 224-227.
71. Pearl J. Causal diagrams for empirical research. Biometrika 1995; 82: 669-710.
72. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. Epidemiology 1999; 10: 37-48.
73. Robins J, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. Epidemiology 2000; 11: 550-60.
74. Robins JM. Data, design, and background knowledge in etiologic inference. Epidemiology 2001; 11: 313-320.
75. Hernan MA, Hernandez-Diaz S, Werler MM et al. Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology. Am J Epidemiol 2002; 155: 176-184.
76. Maldonado G, Greenland S. Estimating causal effects. Int J Epidemiol 2002; 31: 422-429.

77. Vansteelandt S, Goethebeur E. Causal inference with generalised structural mean models. J Royal Stat Society B 2003; 65: 817-835.
78. Frangalis CE, Rubin DB. Principal stratification in causal inference. Biometrics 2002; 58: 21-29.
79. Cox DR, Wermuth N. Multivariate Dependencies. 1996; London: Chapman and Hall.
80. Wermuth N, Cox DR. Statistical dependence and independence. In Encyclopedia of Epidemiological Methods. Gail MH, Benichou J (eds). 2000. John Wiley: New York.
81. Clayton D. Some remarks on interpretation of models and their parameters in epidemiology. International Biometrics Conference, 2002.
82. Hogan JW, Lancaster T. Instrumental variables and inverse probability weighting for causal inference from longitudinal observational studies. Statistical Methods in Medical Research 2004; 13: 17-48.
83. Last JM. A Dictionary of Epidemiology. 2001; IEA: Oxford.

Table 1. Mutually adjusted coefficients obtained from a univariate regression model for offspring birth size*, and from a multivariate regression model for offspring birth size, maternal birth size and maternal adult height in the *Children of the 1950s Study* (N†=1692).

Table 1a		**Univariate model for G2 birth size**	
Explanatory variables	**Units/ Coding**	**Coef.‡**	**95% CI**
G1 birth size	**1 SD**	0.19	0.15, 0.24
G1 adult height	**1 SD (=6.0 cm)**	0.18	0.12, 0.23
G1 parity	**Parous *vs.* nulliparous**	0.26	0.12, 0.40
G0 adult height	**1 SD (=5.6 cm)**	-0.02	-0.07, 0.03
G0 parity	**Parous *vs.* nulliparous**	-0.06	-0.10,-0.03

Table 1b		**Multivariate model for G1 birth size, G1 adult height and G2 birth size**					
		Direct effects on:					
Explanatory		**G2 birth size**		**G1 adult height**		**G1 birth size**	
variables	**Units/ Coding**	**Coef.‡**	**95% CI**	**Coef.‡**	**95% CI**	**Coef.‡**	**95% CI**
G1 birth size	**1 SD**	0.19	0.15, 0.24	0.19	0.15, 0.23	-	-
G1 adult height	**1 SD (=6.0 cm)**	0.18	0.12, 0.23	-	-	-	-
G1 parity	**Parous *vs.* nulliparous**	0.26	0.12, 0.40	-	-	-	-
G0 adult height	**1 SD (=5.6 cm)**	-0.02	-0.07, 0.03	0.49	0.45, 0.53	0.20	0.16, 0.25
G0 parity	**Parous *vs.* nulliparous**	-0.06	-0.10,-0.03	-0.08	-0.11,-0.05	0.07	0.04, 0.11

N: study size; Coef.: estimated regression coefficient; CI: confidence interval

G0: grand-maternal; G1: maternal (index): G2: offspring.

* Birth size is defined as birth weight standardized for gestational age and standardized to have SD=1; this corresponds to around 0.5-0.6 kg in birth weight at most weeks of gestation.

†There were 32 participants with missing values in at least one of the explanatory variables used in these models.

‡Mutually adjusted coefficients.

Table 2. Direct, indirect and total effects estimated from the multivariate model for offspring birth size[a], maternal birth size and maternal adult height in the *Children of the 1950s Study* (N=1692).

		Direct effects		*Indirect effects*		*Total effects*	
Explanatory Variables	**Units/ Coding**	**Coef.**	**95% CI**	**Coef.**	**95% CI**	**Coef.**	**95% CI**
G1 birth size	**1 SD**	0.19	0.15,0.24	0.03	0.02,0.05	0.23	0.18,0.27
G1 adult height	**1 SD (=6.0 cm)**	0.18	0.12,0.23	-	-	0.18	0.12,0.23
G1 parity	**Parous *vs* nulliparous**	0.26	0.12,0.40	-	-	0.26	0.12,0.40
G0 adult height	**1 SD (=5.6 cm)**	-0.02	-0.07,0.03	0.13	0.10,0.16	0.11	0.07,0.16
G0 parity	**Parous *vs* nulliparous**	-0.06	-0.10,-0.03	0.003	-0.01,0.01	-0.06	-0.10,-0.02

N: study size; Coef.: estimated regression coefficient; CI: confidence interval; BW: birth weight; G0: grand-maternal; G1: maternal: G2: offspring

[a] Birth size is defined as birth weight standardized for gestational age and standardized to have SD=1; this corresponds to around 0.5-0.6 kg in birth weight at most weeks of gestation.

Table 3. Univariate linear regression models for leg length at age 43 years on childhood height measures using either absolute childhood height measures or absolute height at age 2 years plus height increments after that age; women in the Medical Research Council National Survey of Health and Development; United Kingdom.

	Univariate linear regression models for adult leg-length *(cm*							
	Model 1		***Model 2***		***Model 3***		***Model 4***	
	(N=1196)		**(N=1050)**		**(N=1001)**		**(N=920)**	
	Coef.	**95% CI**	**Coef.**	**95% CI**	**Coef.**	**95% CI**	**Coef.**	**95%**
Table 3a								
***Height** (cm) **at age** (years)*:								
2	0.08	0.03, 0.14	-0.02	-0.08, 0.05	-0.02	-0.08, 0.04	-0.03	-0.09,
4	0.37	0.32, 0.43	0.10	0.03, 0.18	0.05	-0.03, 0.13	0.04	-0.05,
6	-	-	0.38	0.31, 0.46	0.14	0.03, 0.24	0.15	0.04,
7	-	-	-	-	0.31	0.21, 0.41	0.29	0.17,
11	-	-	-	-	-	-	0.03	-0.03,
15*	-	-	-	-	-	-	-	-
Table 3b								
***Height** (cm) **at age** (years)*								
2	0.45	0.40, 0.51	0.47	0.41, 0.53	0.48	0.42, 0.54	0.48	0.42,
***Height increment** (cm) **between ages** (years)*:								
2 and 4	0.37	0.32, 0.43	0.49	0.43, 0.54	0.50	0.44, 0.55	0.51	0.45,
4 and 6	-	-	0.38	0.31, 0.46	0.44	0.37, 0.52	0.47	0.39,
6 and 7	-	-	-	-	0.31	0.21, 0.41	0.32	0.21,
7 and 11	-	-	-	-	-	-	0.03	-0.03,
11 and 15*	-	-	-	-	-	-	-	-

Coef.: estimated regression coefficient; CI: confidence interval

* Height was measured between 14 and 15 years.

Table 4. Univariate linear regression models for leg length at age 43 years on a selection of childhood height measures specified either as absolute measures or absolute height at age 2 years plus height increments or absolute height at age 2 years plus height velocities*; women in the Medical Research Council National Survey of Health and Development; United Kingdom.

			Univariate linear regression for adult leg-length (N=794)	
	Mean	**SD**	**Coef.**	**95% CI**
Table 4a				
***Height** (cm) **at age** (years)*:				
2	84.7	4.8	-0.06	-0.13, 0.00
7	119.5	5.7	0.19	0.11, 0.28
11	140.9	7.2	-0.10	-0.16. -0.03
15 †	158.5	6.3	0.45	0.38, 0.52
Table 4b				
Height** (cm) **at 2 years	84.7	4.8	0.48	0.43, 0.54
***Height increment** (cm) **between ages** (years*:				
2 and 7	34.8	5.2	0.55	0.50, 0.60
7 and 11	21.4	4.0	0.36	0.28, 0.43
11 and 15†	17.6	4.5	0.45	0.38, 0.52
Table 4c				
Height** (cm) **at 2 years	84.68	4.80	0.48	0.43, 0.54
Height velocity (cm/years) **between ages** (years)*:				
2 and 7	6.95	1.04	2.74	2.48, 3.01
7 and 11	5.34	0.99	1.43	1.13, 1.73
11 and 15 †	4.39	1.13	1.81	1.54, 2.08

Coef.: estimated regression coefficient; CI: confidence interval

*Velocity is defined as height increment divided by ages difference.

† Height was measured between 14 and 15 years.

Table 5. Joint estimation of the measurement and structural models described in Figure 6 obtained using only complete records or all records with at least one of the measures and obtained via EM-ML assuming MAR; women in the Medical Research Council National Survey of Health and Development; United Kingdom.

	Complete records analysis *(N=794)*			***Under MAR assumption*** *(N=2349)*		
	Coef.	**Expected mean**	**95% CI**	**Coef.**	**Expected Mean**	**95% CI**
Measurement model for latent growth						
Height *(cm)* ***at 2 years***		87.3	87.0,87.6		87.3	87.0, 87.5
Height velocity *(cm/year)* ***between ages*** *(years)**:***						
2 and 7		6.59	6.53, 6.66		6.51	6.45, 6.57
7 and 11		5.26	5.19, 5.34		5.27	5.22, 5.32
11 and 15 *		4.37	4.31, 4.43		4.41	4.36, 4.44
Structural model for leg length						
Height *(cm)* ***at 2 years***	0.48		0.39, 0.56	0.45		0.38, 0.52
Height velocity *(cm/year)* ***between ages*** *(years)**:***						
2 and 7	2.29		1.67, 2.91	2.49		1.98, 3.00
7 and 11	1.79		1.34, 2.25	1.96		1.54, 2.31
11 and 15 *	1.96		1.47, 2.45	2.31		1.84, 2.77

N: study size; Coef.: estimated regression coefficient; CI: confidence interval; EM: expectation-maximization algorithm; ML: maximum likelihood; MAR: missing at random.

* Height was measured between 14 and 15 years.

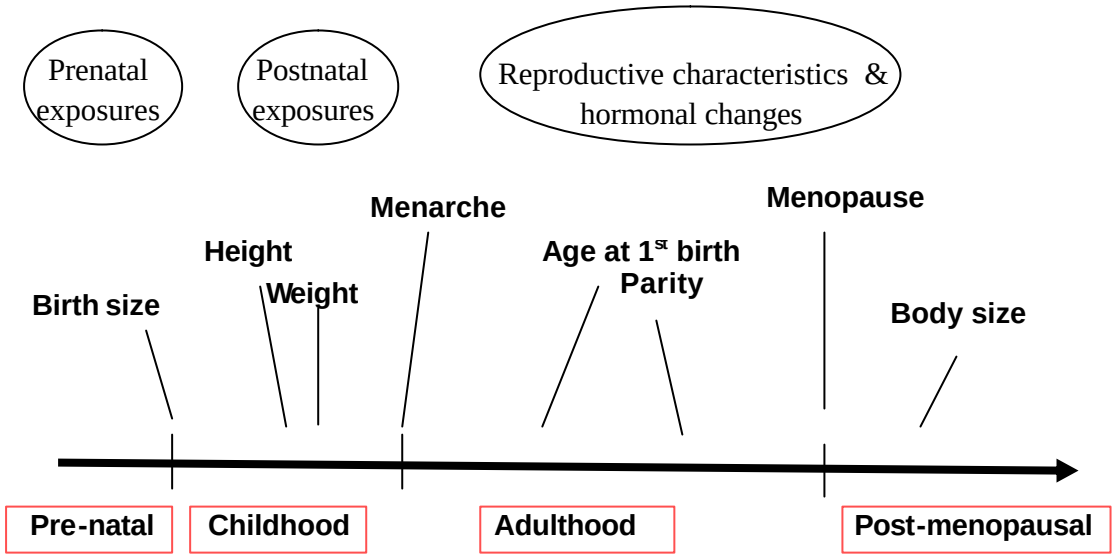
Table 6 Characteristics of women with/without complete childhood height and adult leg-length data; women in the Medical Research Council National Survey of Health and Development; United Kingdom.

	With complete records			Without complete records		
	N	**Mean**	**SD**	**N**	**Mean**	**SD**
***Height** (cm) **at age** (years)*:						
2	794	85.08	4.40	1079	84.38	5.05
7	794	119.96	5.39	1199	119.18	5.88
11	794	141.36	6.99	1119	140.49	7.24
15 †	794	158.81	6.14	932	158.15	6.42
Height velocity (cm/years) **between ages** (years)*:						
2 and 7	794	6.98	1.00	839	6.93	1.08
7 and 11	794	5.35	1.01	1013	5.34	0.98
11 and 15 †	794	4.36	1.13	841	4.42	1.14
Leg-length	794	75.45	4.57	814	75.66	4.84

* Velocity is defined as height increment divided by ages difference.

† Height was measured between 14 and 15 years.

FIGURE 1. A simplistic time line representation of conceptual (in the ovals) and observable risk factors for breast cancer.



THE LIFE COURSE

Figure 2. Hypothetical life course plot indicating that change in the explanatory variable between age 4 and 6 years has a positive effect on the outcome

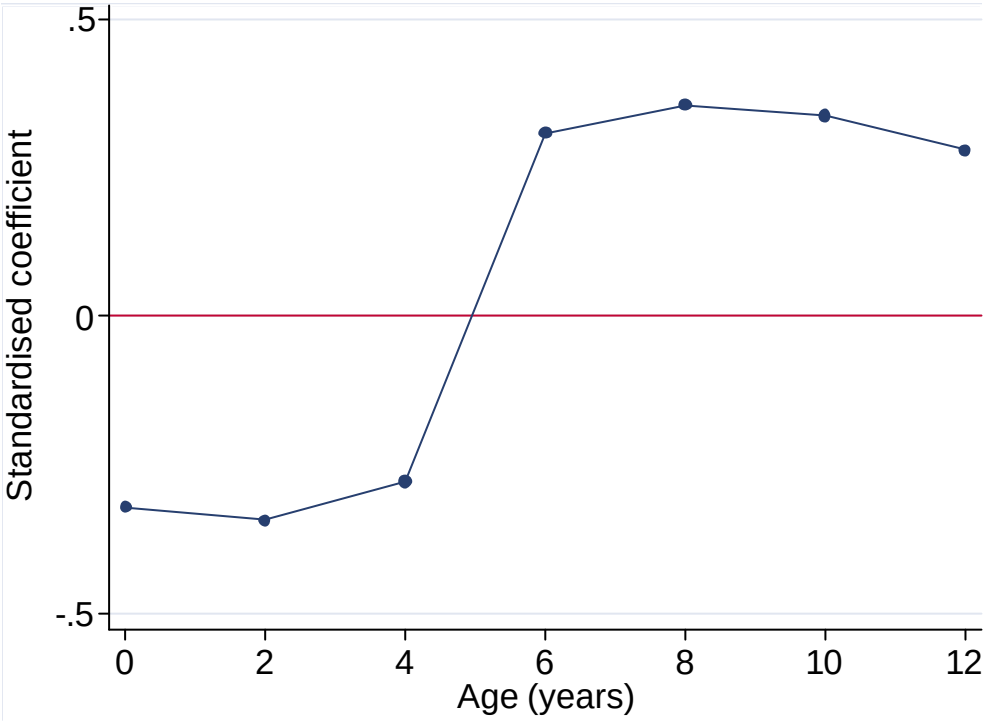
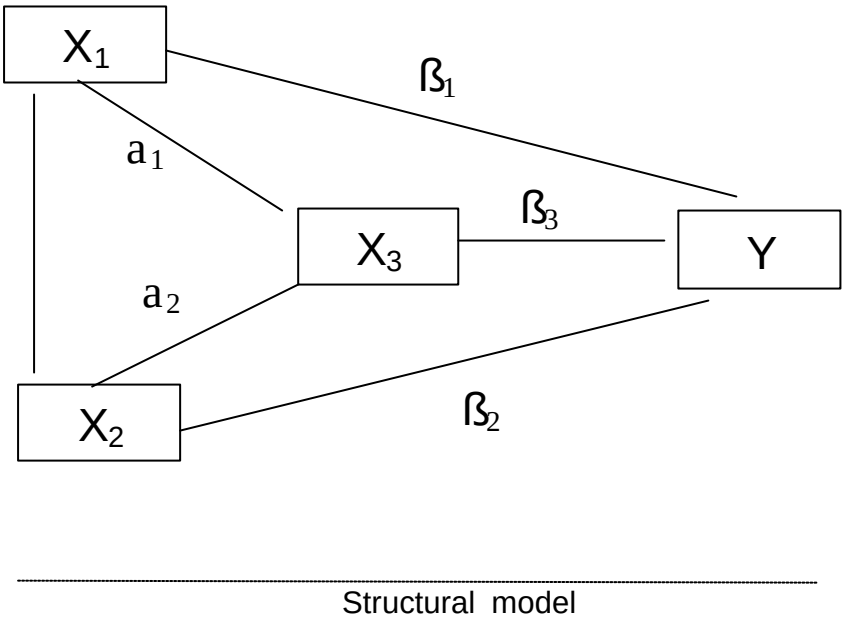
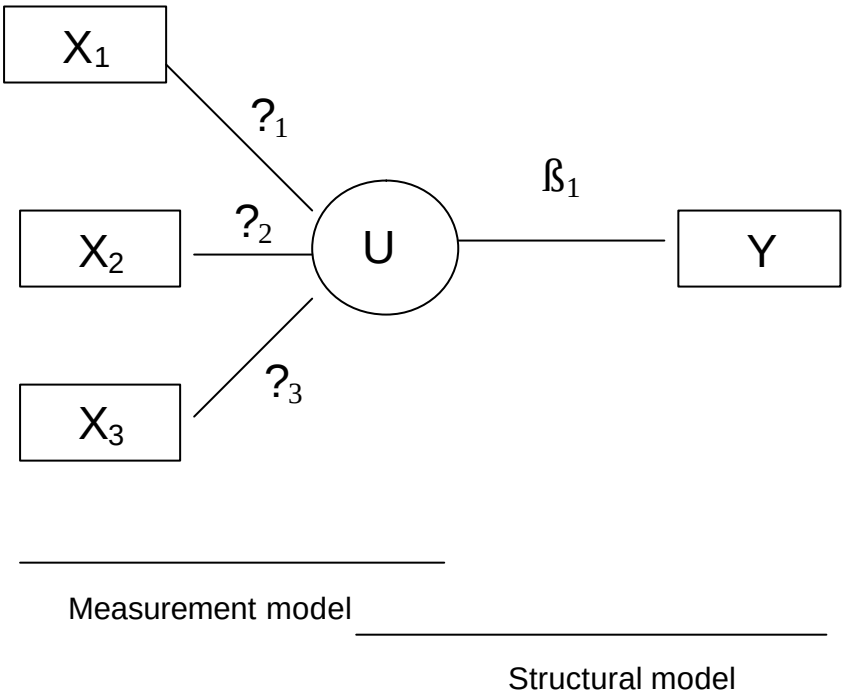


Figure 3a – Example of path diagram* for one distal outcome (Y), one intermediate outcome (X_3) and two background variables (X_1, X_2).



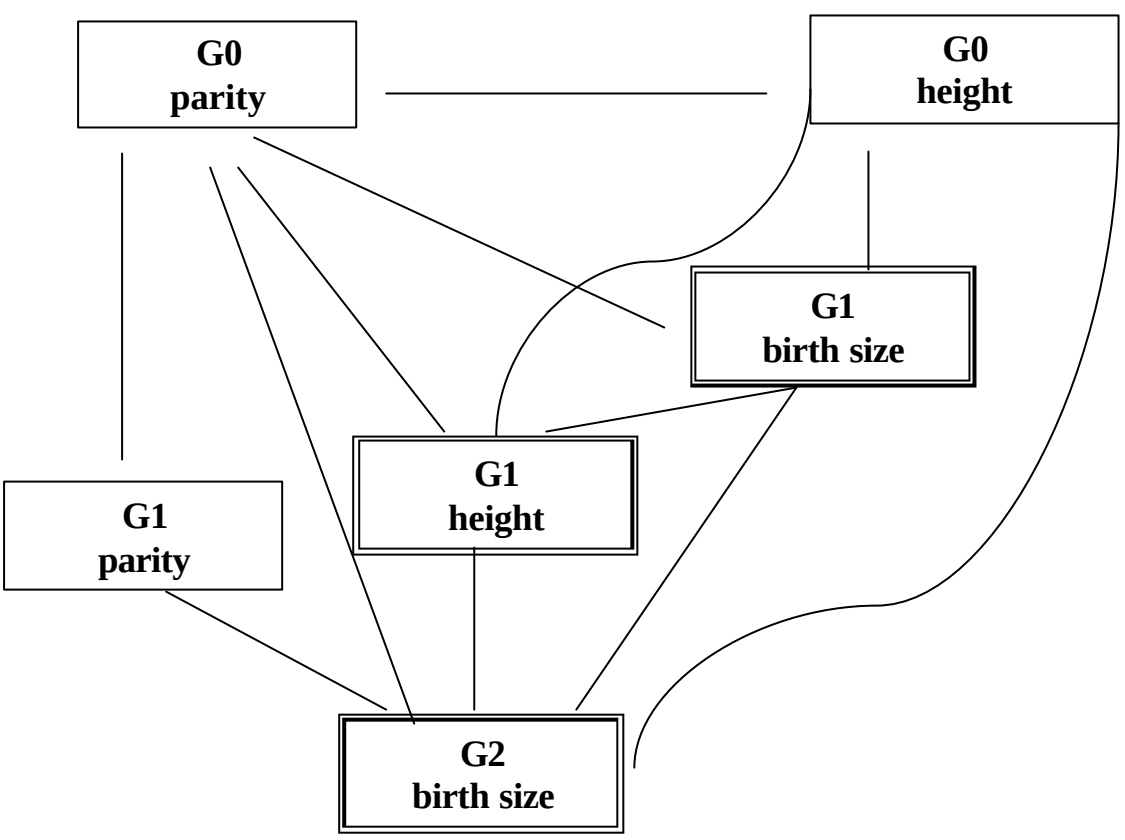
* Arrows depicting random variation for each variable are omitted for simplicity.

Figure 3b – Example of path diagram* for one distal outcome (Y) and a latent variable (U) measured by three proxy variables (X_1, X_2, X_3)



* Arrows depicting random variation for each variable are omitted for simplicity. Boxes are used to represent proxy variables, circles for latent variables.

Figure 4. Simplified path diagram* for intergenerational influences on birth size



G0: grandmothers; G1: mothers (index); G2: offspring

* Arrows depicting random variation for each variable are omitted for simplicity; double lined boxes are the dependent variables.

Figure 5. Life course plot of the regression coefficients in the model of leg length on standardised height measures from age 2 to age 15 years; the Medical Research Council National Survey of Health and Development; United Kingdom; N=791.

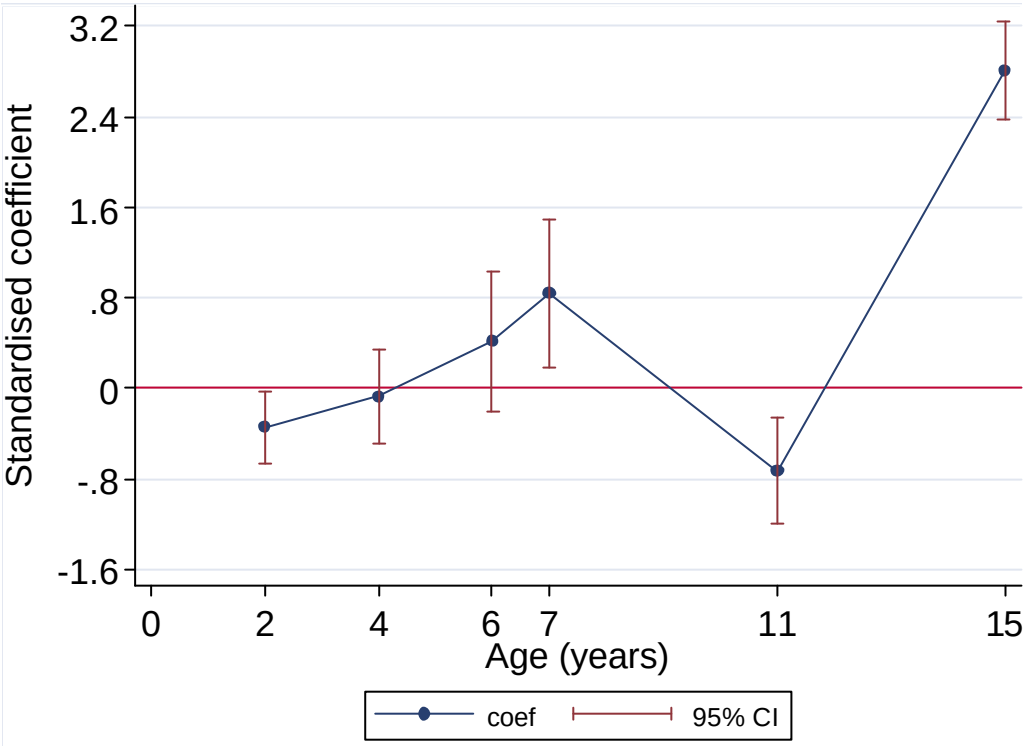


Figure 6. Hypothetical piecewise linear model for childhood growth

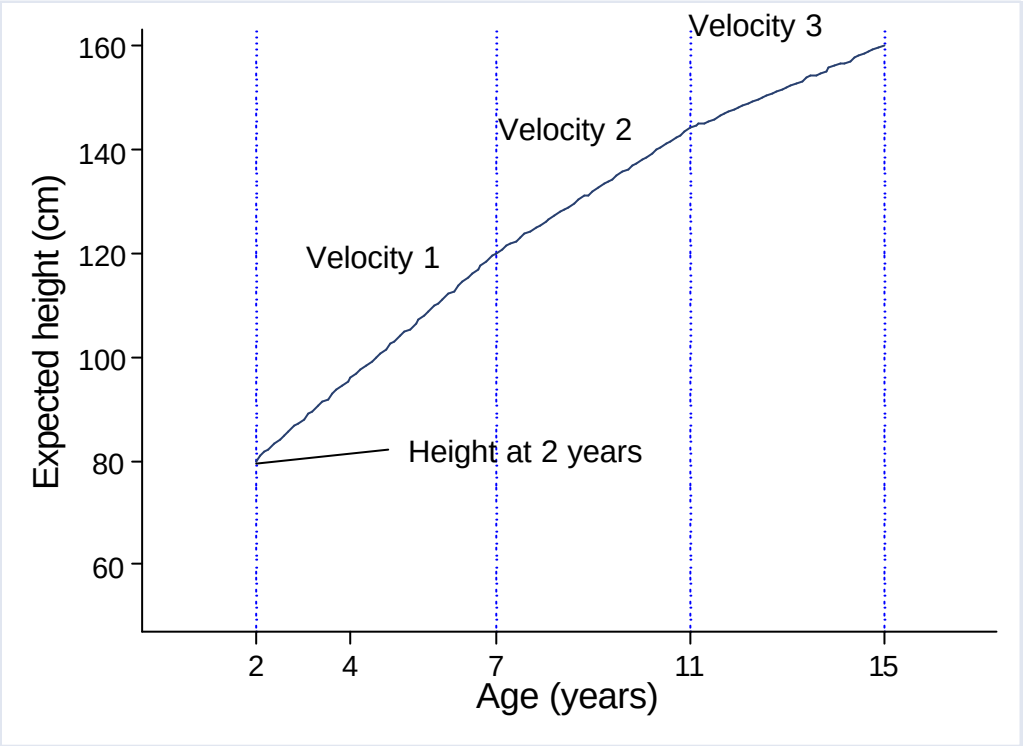
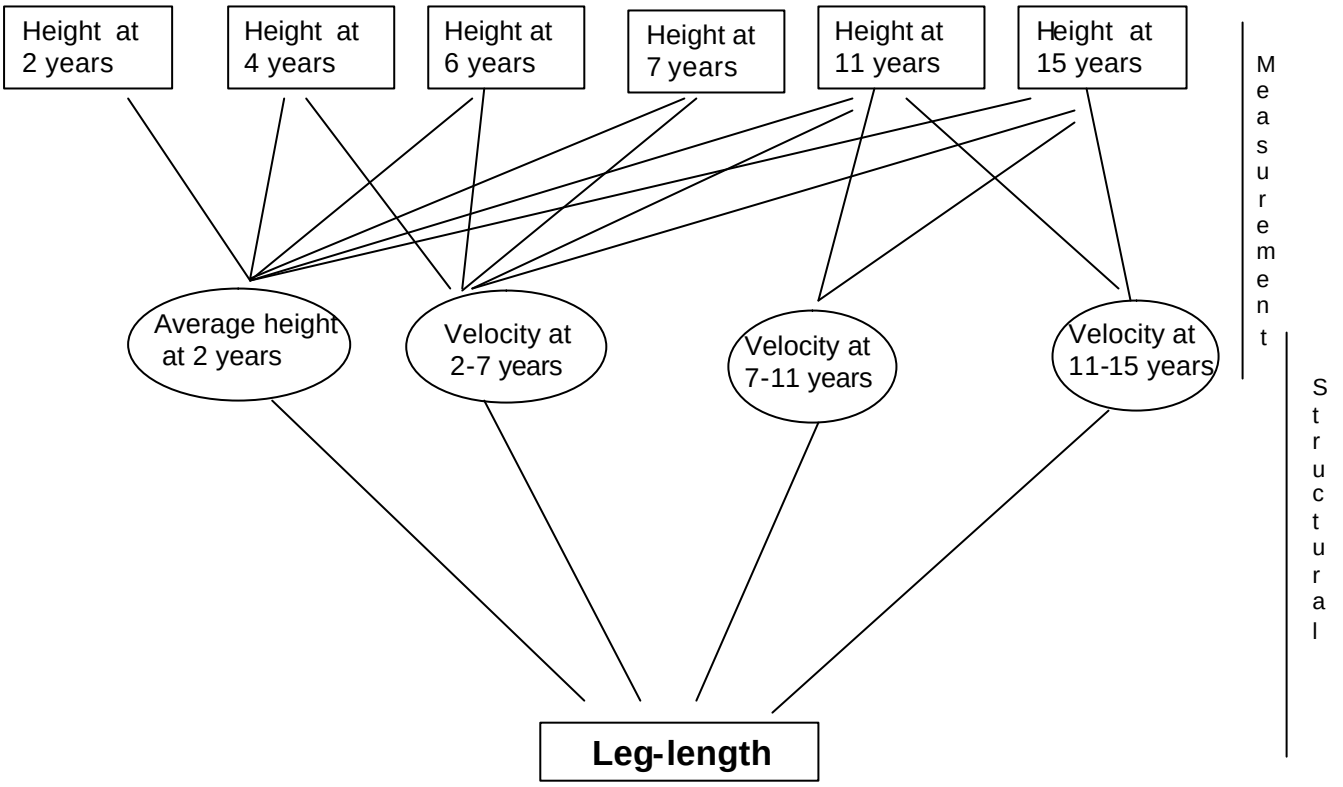


Figure 7. Simplified path diagram* of the relationship between the height measurements taken at ages 2, 4, 7, 11, and 15 years and adult leg length.



* Arrows depicting random variation for each variable are omitted for simplicity. Boxes are used to represent proxy variables, circles for latent variables.