OPEN

Model Selection of the Effect of Binary Exposures over the Life Course

Andrew D. A. C. Smith,[a,b] *Jon Heron,*[a] *Gita Mishra,*[c] *Mark S. Gilthorpe,*[d] *Yoav Ben-Shlomo,*[a] *and Kate Tilling*[a,b]

Abstract: Epidemiologists are often interested in examining the effect on a later-life outcome of an exposure measured repeatedly over the life course. When different hypotheses for this effect are proposed by competing theories, it is important to identify those most supported by observed data as a first step toward estimating causal associations. One method is to compare goodness-of-fit of hypothesized models with a saturated model, but it is unclear how to judge the "best" out of two hypothesized models that both pass criteria for a good fit. We developed a new method using the least absolute shrinkage and selection operator to identify which of a small set of hypothesized models explains most of the observed outcome variation. We analyzed a cohort study with repeated measures of socioeconomic position (exposure) through childhood, early- and mid-adulthood, and body mass index (outcome) measured in mid-adulthood. We confirmed previous findings regarding support or lack of support for the following hypotheses: accumulation (number of times exposed), three critical periods (only exposure in childhood, early- or mid-adulthood), and social mobility (transition from low to high socioeconomic position). Simulations showed that our least absolute shrinkage and selection operator approach identified the most suitable hypothesized model with high probability in moderately sized samples, but with lower probability for hypotheses involving change in exposure or highly correlated exposures. Identifying a single, simple hypothesis that represents the specified knowledge of the life course association allows more precise definition of the causal effect of interest.

(*Epidemiology* 2015;26: 719–726)

From the [a]School of Social and Community Medicine, University of Bristol, Bristol, United Kingdom; [b]MRC Integrative Epidemiology Unit at the University of Bristol, Bristol, United Kingdom; [c]School of Population Health, University of Queensland, St Lucia, QLD, Australia; and [d]Division of Epidemiology and Biostatistics, School of Medicine, University of Leeds, Leeds, United Kingdom.

AS, JH, MSG, and KT were supported by the Medical Research Council (Grant Number G1000726). AS and KT work in a Unit that receives funding from the UK Medical Research Council and the University of Bristol (MC_UU_12013/9).

The authors report no conflicts of interest.

SDC Supplemental digital content is available through direct URL citations in the HTML and PDF versions of this article (www.epidem.com). This content is not peer-reviewed or copy-edited; it is the sole responsibility of the authors.

Correspondence: Andrew D. A. C. Smith, Oakfield House, Oakfield Grove, Bristol BS8 2BN, United Kingdom. E-mail: Andrew.D.Smith@bristol.ac.uk.

ISSN: 1044-3983/15/2605-0719
DOI: 10.1097/EDE.0000000000000348

Medical research over the past two decades has examined fetal and early life antecedents of disease, and their interaction with other exposures throughout the life course to influence later-life conditions.[1] Several hypothetical relations between repeated covariate measures (e.g., repeated measures of socioeconomic position in childhood) and a subsequent outcome (e.g., adult blood pressure) can be proposed, based on theoretical models or mechanisms of action.[2,3] For example, a hypothesized "critical period" in early childhood specifies that socioeconomic position during this period has lasting effects on blood pressure. An alternative hypothesis is of an "accumulation" of risk across the life course (i.e., that adverse social circumstances at any time increases subsequent risk of high blood pressure). The hypotheses examined should inform the analytic methods used.[4–6] The first step in estimating a causal association is to specify knowledge about the system—in this case the life course—being studied.[7] This knowledge will be incomplete unless the hypothesized relation between exposures and outcome has been investigated. The initial investigation may be thought of as exploratory: determining the most likely relation, while later analyses may be thought of as confirmatory: verifying the hypothesized relation and checking that no other relations are present.

There has been growing interest in a structured approach to life course hypotheses, in which a closed set of hypotheses is proposed and tests conducted to identify best-fitting hypotheses.[8] One such approach uses an F test to compare a saturated model with hypothesized models concerning the association between binary exposure variables, measured over the life course, and an outcome.[9] This method has been used in several studies with continuous outcomes[10–16] and adapted for binary outcomes.[17–19] A hypothesis may be thought of as supported by observed data if the F test yields a P value above a certain threshold, although a large P value cannot be considered to "prove" the hypothesis. If more than one hypothesis passes the threshold, the hypothesis that renders the largest P value, or smallest Akaike information criterion,[20,21] may be selected. The performance of these methods in the life course setting has not been formally assessed.

We describe an alternative model selection strategy that identifies which hypothesis, selected from an a priori-compiled set of hypotheses, explains the most variation in

the outcome. We illustrate how the proposed method can be used in both exploratory and confirmatory studies. The performance is contrasted with the structured F test approach in data from a previously published example,[9] and also through simulation.

METHODS

The proposed strategy involves selecting from an a priori-compiled set of potential hypotheses describing the association between exposure over the life course and outcome. Each hypothesis is encoded into one or more variables, which are then all included in a regression model, and the subset of variables that explains the greatest proportion of the outcome variation is selected. The number of hypotheses is not limited; any hypothesis may be examined provided there are enough exposure measurements to identify it. Several variations of similar hypothesis may be considered (e.g., a set of critical period hypotheses covering a range of possibly overlapping periods). The choice of hypotheses to examine may be informed by knowledge of causal mechanisms, perhaps using a directed acyclic graph. Examples of typical hypotheses for binary exposures are described below.

An *accumulation* hypothesis states that there is a linear association between the outcome and the cumulative sum of the exposure over the life course.

A *critical period* hypothesis states that only exposure during one period is associated with the outcome. Under a *sensitive period* hypothesis, the outcome is associated with the amount of exposure, as in the accumulation hypothesis, but the association is stronger in a particular period.[21]

A *mobility* hypothesis states that the outcome is associated with changes in the exposure over time. The simplest mobility hypotheses relate the outcome only to unidirectional changes. A more complex mobility hypothesis may relate the outcome to bidirectional changes[9,22] (e.g., a positive association with increased exposure and a negative association with decreased exposure). This would in general enhance the plausibility of a causal association. The related *interaction* hypothesis states that the outcome is associated with the exposure in a particular period, but that this association is altered by the exposure in a different period.

Encoding of Variables

Each hypothesis is encoded as a variable that is proportional to the hypothesized outcome as the exposure varies. Simpler hypotheses may be encoded by a single variable; more complex hypotheses need multiple variables. Below, we give details of the single variables that encode the simple hypotheses discussed above. We assume a set of m repeated binary measures of exposure $X_1,...,X_m$.

The *accumulation* hypothesis is encoded by the variable $A = X_1 + \cdots + X_m$.

If there is only one measurement occasion during a hypothesized *critical period* then only that measurement will be associated with the outcome. A hypothesis of a critical period at the jth measurement occasion is encoded by the variable $C_j = X_j$. If there are several measurement occasions during the critical period (e.g., $X_j,...,X_k$), then the critical period hypothesis may be encoded by $C_{jk} = X_j + \cdots + X_k$, i.e., as accumulation within the critical period.

Under the simplest *mobility* hypothesis, the outcome varies with a unidirectional change in the exposure. A mobility hypothesis between the jth and kth measurement occasions may be encoded by $M^+_{jk} = (1 - X_j)X_k$ if it is hypothesized that a positive change from j to k is associated with the outcome, or by $M^-_{jk} = X_j(1 - X_k)$ if it is hypothesized that a negative change from j to k is associated with the outcome.

Some hypotheses require more than one variable to encode, and can therefore be thought of as compound hypotheses. All of these can be encoded by combinations of variables encoding simple hypotheses; some quite complex hypotheses can be encoded with two variables. The combinations of variables that encode compound hypotheses are described below, with further details provided in eAppendix A (http://links.lww.com/EDE/A940).

A *sensitive period* hypothesis can be encoded by the combination of the accumulation variable and the relevant critical period variable.

Two simple mobility variables can, together, encode a more complex mobility hypothesis. For instance, mobility hypotheses may combine a variable encoding positive change with a variable encoding negative change, or variables encoding change at different pairs of measurement occasions over the life course.[9] An *interaction* hypothesis can be encoded by combining critical period and mobility variables. For example, combining the critical period variable C_j with mobility variable M^+_{jk} encodes a hypothesis that the outcome is associated with the exposure measurement at occasion j, but this association is modified by the exposure measurement at occasion k.

Choosing Hypotheses Most Strongly Supported by Observed Data

After encoding potential hypotheses, our approach is to examine the association between all encoded variables and the outcome, and select only those encoded variable combinations that have the strongest association with the outcome. Since there are potentially more variables than available degrees of freedom it is inappropriate to put all encoded hypotheses into a linear model and choose the variable(s) with the largest parameter estimates. Instead, we propose placing an absolute value penalty on parameter estimates, whereby unimportant variables have their estimated association shrunk to zero. Hence, the resulting fit will provide the fullest explanation of the observed data from the fewest parameters. The Least Absolute Shrinkage and Selection Operator (lasso),[23] which minimizes the residual sum of squares plus an absolute value penalty, provides a suitable method, but requires selection of a smoothing parameter. This can be simplified by implementing

the Least Angle Regression (LARS) approach to the lasso,[24] which provides lasso estimates for all smoothing parameter values and indicates the best lasso fit for each number of selected variables, reducing the problem to that of choosing the number of variables. The lasso is a constrained version of linear regression and may be used whenever the assumptions of linear regression are satisfied. The LARS algorithm first selects the variable with the strongest association with the outcome,[25] hence this approach will always select first the hypothesis, or component of a compound hypothesis, that offers the strongest explanation for the observed data. Using an absolute value penalty causes subsequent variables to be added in order of strength of association with outcome variation.[26] The overall hypothesized view of the data is thus built from the most relevant simple hypotheses or components of compound hypotheses.

When little is known regarding the association between the exposures and outcome over the life course, the structured approach can be used to suggest the most likely of an a priori-defined set of hypothesized associations. This set may extend to several hypotheses if we have little a priori information about which are likely. In this exploratory setting, the choice of hypothesis is somewhat subjective and we require a method for choosing how many variables to include in our selected hypothesis. We use an elbow plot—a plot of the proportion of outcome variation explained by the lasso fit (the R^2 value) against number of variables selected at each stage in the LARS procedure. The "elbow"—a sharp concave bend at which adding more variables does not substantially increase the R^2 value—is used to choose the number of variables. Provided enough variables are included in the procedure, the elbow plot will show how the lasso selections approach a saturated model. It is useful to see the R^2 value for a saturated model to check whether there is any association between the outcome and all exposure measurements over the life course. An alternative method for selecting the number of variables is provided by the lasso covariance hypothesis test.[27] At each stage of the LARS procedure, this tests the null hypothesis that adding the next variable does not improve the R^2 value. The covariance hypothesis test accounts for the fact that the next variable will have the greatest association out of the variables not already selected. An alternative option, a nested F test to discriminate between simple and complex hypotheses,[22] may be biased by the selection of the simpler model due to its greater association. In the exploratory setting, it may be necessary to reject more compound hypotheses in favor of simpler ones to maintain plausibility.

We may have a firmer idea of the nature of the life course association between exposures and outcome, and perhaps some causal information. In this setting, we might specify quite a small number of possible hypotheses a priori, and rather than choose the number of variables in our selected hypothesis we might simply choose the first variable or hypothesis selected by the LARS algorithm, to confirm our causal assumptions.

EXAMPLE

The proposed approach, in the exploratory setting, is illustrated using data on socioeconomic position and body mass index (BMI) from a cohort of 2,192 men and women, with binary measurements of the exposure, socioeconomic position, at ages 4, 26, and 43 years, and a continuous measurement of the outcome, BMI, at age 53 years. These data were previously used to illustrate the structured approach, and full details are given in the original study.[9] Issues such as confounding and measurement error were not the focus of the original study and are therefore ignored here for sake of simplicity and comparison with the alternative structured method. Figure 1 shows a possible directed acyclic graph for this example. While there are many potential confounders of the life course association between exposure and outcome, which in this example are unmeasured (as they were in the original study), the focus in this exploratory setting is to identify likely life course associations between exposure and outcome, whether or not they are confounded. We considered the set of six hypotheses that have been previously proposed for this data: three critical period hypotheses corresponding to the three exposure measurement occasions, two mobility hypotheses concerning change between adjacent measurement occasions, and an accumulation hypothesis.[9]

Encoding of Variables

Each hypothesis was encoded based on the binary exposure measurements X_1, X_2, and X_3 at the three time points, where zero represents a manual socioeconomic position and one a nonmanual socioeconomic position. The simple hypotheses, requiring one variable each, were the three critical period hypotheses, encoded by C_1, C_2, and C_3, and the accumulation hypothesis, encoded by $A = X_1 + X_2 + X_3$. The two mobility hypotheses required two variables each, with M_{12}^{+} and M_{12}^{-} encoding mobility between ages 4 and 26 years, and M_{23}^{+} and M_{23}^{-} encoding mobility between ages 26 and 43 years.

Use of Elbow Plot

Figure 2 shows the elbow plot for men. The variance explained by the model with the greatest number of variables was 1.7%, which approached the 2% explained by the saturated model; hence the maximum R^2 value on the plot is close to the maximum R^2 value achievable. There is a clear elbow where one variable is selected; adding additional variables did not considerably improve the R^2 value. In addition, the P value for adding a second variable was 0.90, indicating no evidence that one variable was insufficient. The first variable selected encoded the hypothesis of an age 4 critical period; choosing this elbow point identified this hypothesis as offering the best explanation for the observed data in men.

Figure 3 shows the elbow plot for women. The maximum R^2 value on the plot is 3.8%; that of the saturated model

was 3.9%. Therefore, socioeconomic position explained a greater proportion of the BMI variation in women than men. The position of an elbow point was less clear for women. The first variable selected encoded the accumulation hypothesis; the *P* value for adding a second variable was 0.52. It could thus be concluded that the accumulation hypothesis is the primary explanation for the observed data in women. The plot might also be considered to have an elbow at three variables, selecting variables that encoded the accumulation, childhood critical period, and adult nonmanual to manual mobility variables. Our interpretation of this is a hypothesis of two sensitive periods: sensitivity (to manual socioeconomic position) in childhood and sensitivity to change (from nonmanual to manual socioeconomic position) in adulthood.

In the exploratory setting, further study would be required to clarify these hypotheses and examine causal mechanisms.

SIMULATION STUDY

To investigate how frequently the LARS algorithm selects the correct hypothesis in a confirmatory setting, we simulated data with two and three repeated binary exposures. The performance indicator is the selection probability: the proportion of simulations where the correct hypothesis is identified.

Two Exposure Measurements

Two exposure measurements were simulated as binary random variables, being zero or one with equal probability (see eAppendix B for details; http://links.lww.com/EDE/A940). We considered the situation in which the outcome is known to be associated with change in the exposure, but it is yet to be confirmed whether a mobility or interaction hypothesis defines the true association. We simulated mobility and interaction models, varying the correlation, ρ, between exposure variables, and the residual variance, σ^2. In each simulation, we selected from six proposed compound hypotheses (four interaction hypotheses, full mobility, and an additive model).

We compared three approaches for identifying the hypotheses offering the best explanation for the simulated data. The first was the LARS algorithm for the lasso, which was considered to have identified the correct hypothesis if the first two selected variables encoded that hypothesis. The second approach chose the hypothesized model with the largest *F* test *P* value when compared with the saturated model,[9] and the third approach selected the hypothesized model yielding the smallest Akaike information criterion of those models with an *F* test *P* value not less than 0.05.[21]

We varied the sample size from 400, which might represent a subset of a study, to 2,500, which might represent a moderate-sized study. We ran 500 simulations for each combination of residual variance, correlation, model, and sample size. With 500 simulations, the 95% confidence interval for the selection probability will have a radius of less than 2% for a selection probability of 95%, and less than 5% for a selection probability of 50%.

Table 1 shows the selection probabilities, in simulation, of the three methods. The selection probability of the LARS algorithm is very good in situations with low residual variance or large sample size: it was at least 83.6% when the R^2 value was at least $100/n$. This approach always outperformed the alternative methods, except for a difference of 0.4% in one situation with the largest residual variance and smallest sample size.

Three Exposure Measurements

This hypothetical example considered that prior knowledge provided evidence for either a critical or a sensitive period; the aim being to confirm the correct association using new data. Three measurements were simulated as binary variables as before, with adjacent measurements having correlations of ρ, and the first and last measurements having correlation ρ^2. The models used to generate the outcome were: an early critical period, an accumulation model, and an early sensitive period model. The simple hypotheses that the LARS algorithm was allowed to choose from were three critical period hypotheses and an accumulation hypothesis. The LARS algorithm was considered to have chosen a simple hypothesis if it first selected the variable encoding that hypothesis, and the covariance test *P* value for including another variable was not less than 0.05. The LARS algorithm was considered to have identified a compound hypothesis if the first two selected variables encoded that hypothesis and the *P* value for including the second variable was less than 0.05. While we do not necessarily advocate using *P* value thresholds to select hypotheses, this allowed comparisons with other approaches. The other two approaches used the *F* test and Akaike information criterion as before, choosing between an accumulation model, three critical period models and three sensitive period models.

There was some evidence that selection probabilities decreased as the exposure correlation increased (Table 2). The selection probability of the LARS algorithm was very good with low residual variance or large sample size: it was at least 90.2%, and higher than that of other methods, when the R^2 value was at least $100/n$. The only exception to this was in sensitive period simulations with strong exposure correlation, where there was less distinction between accumulation and critical period hypotheses. The alternative methods had better selection probabilities in compound models than simple models; it appears that Akaike information criterion or *F* test selection is more likely to select compound hypotheses over simple ones, regardless of the true underlying model.

Confidence Intervals

We repeated the simulation experiment with three exposure measurements, testing a null model in place of a sensitive period model. In each simulation, we calculated the usual

TABLE 1. Percentage of 500 Simulations in Which the Correct Model Was Identified, in Simulations with Two Binary Exposure Measurements, by Three Different Structured Approaches

		n = 400			n = 1,000			n = 2,500		
$\sigma^2 = 1$ ($R^2 = 0.50$)	ρ	0	0.4	0.8	0	0.4	0.8	0	0.4	0.8
Mobility	LARS	99.4	100	99.8	100	100	100	100	100	100
	F test	51.6	52.0	51.0	50.8	51.8	52.2	53.6	50.2	49.2
	AIC and *F* test	51.2	52.0	51.0	50.8	51.8	51.8	53.6	50.2	48.8
Interaction	LARS	100	100	100	100	100	100	100	100	100
	F test	100	100	97.4	100	100	100	100	100	100
	AIC and *F* test	94.6	94.2	95.6	97.8	96.4	94.6	95.8	96.8	94.2
		n = 400			n = 1,000			n = 2,500		
$\sigma^2 = 9$ ($R^2 = 0.10$)	ρ	0	0.4	0.8	0	0.4	0.8	0	0.4	0.8
Mobility	LARS	62.8	70.8	61.0	86.6	95.8	87.4	98.2	100	98.2
	F test	49.4	48.8	39.8	50.8	51.8	49.2	53.6	50.2	49.2
	AIC and *F* test	49.4	48.8	39.8	50.8	51.8	49.2	53.6	50.2	48.8
Interaction	LARS	97.2	97.2	84.2	100	100	97.2	100	100	100
	F test	80.6	75.0	50.0	96.6	95.0	76.8	100	100	94.6
	AIC and *F* test	80.0	75.0	50.0	95.4	94.2	76.8	95.8	96.8	92.8
		n = 400			n = 1,000			n = 2,500		
$\sigma^2 = 24$ ($R^2 = 0.04$)	ρ	0	0.4	0.8	0	0.4	0.8	0	0.4	0.8
Mobility	LARS	38.4	46.2	39.2	57.0	69.8	60.8	83.6	94.0	94.0
	F test	38.8	37.6	27.4	48.0	48.6	37.2	53.6	49.8	46.2
	AIC and *F* test	38.8	37.6	27.4	48.0	48.6	37.2	53.6	49.8	46.2
Interaction	LARS	82.2	80.0	62.4	96.4	97.2	84.2	100	100	97.0
	F test	49.8	45.6	31.4	78.4	73.6	53.6	95.6	93.4	72.8
	AIC and *F* test	49.8	45.6	31.4	78.2	73.6	53.6	93.6	92.0	72.6

AIC indicates Akaike information criterion; LARS, least angle regression.

95% confidence interval for the regression parameter in the hypothesized model with the largest *F* test *P* value (coincidentally the model with smallest Akaike information criterion) when compared with the saturated model. We also calculated an adjusted confidence interval based on the covariance test for the lasso (see eAppendix B for details; http://links.lww.com/EDE/A940). Table 3 shows the coverage of these confidence intervals. In the null model, the coverage of the usual confidence intervals was always less than 95%, showing that the *F* test or Akaike information criterion approach generates bias due to the fact that they consider the largest observed association to be selected at random. However, the adjusted confidence intervals have coverage between 92.2% and 96.2% in the null model, confirming that the covariance test corrects for selection of the variable with greatest association.

DISCUSSION

A causal life course association between exposure and outcome cannot be estimated without identifying knowledge of the system being studied.[7] This can be achieved by assessing prespecified competing hypotheses regarding that system and the life course association. We have described a strategy for this, which involves encoding a set of hypotheses as covariates, and then using the LARS procedure for the lasso to identify the most appropriate covariate subset that accounts for the outcome variation. Variable selection is aided visually with an elbow plot, or guided by a hypothesis test. We showed that, for one example dataset, the LARS procedure identified the same hypotheses as earlier research using a structured approach.[9] Furthermore, simulation showed, for reasonably large sample sizes, the LARS algorithm, even when combined with a naive *P* value threshold, effectively identified the correct hypotheses. Alternative methods, based on *F* tests and Akaike information criterion, did not identify the correct hypotheses as often and were more likely to favor compound hypotheses over simple ones.

Our proposed approach is part of the process toward estimation of causal effects. The set of hypotheses proposed a priori can be chosen using previous knowledge and theory regarding the plausibility of various hypotheses. We have demonstrated techniques for choosing the best-fitting of those hypotheses, which can then be further investigated both for

TABLE 2. Percentage of 500 Simulations in Which the Correct Model Was Identified, in Simulations with Three Binary Exposure Measurements, by Three Different Structured Approaches

		n = 400			n = 1,000			n = 2,500		
$\sigma^2 = 1$ ($R^2 = 0.50$)	ρ	0	0.4	0.8	0	0.4	0.8	0	0.4	0.8
Critical period	LARS	97.2	97.0	97.2	98.2	97.4	97.2	96.6	96.8	97.8
	F test	67.2	66.2	67.8	68.4	66.4	68.2	66.4	66.4	64.0
	AIC and *F* test	83.6	82.2	79.6	81.4	81.8	83.2	83.6	81.4	79.0
Accumulation	LARS	93.6	92.8	92.4	94.4	92.4	90.2	91.6	93.0	91.2
	F test	40.8	37.4	38.4	40.0	38.4	33.8	38.4	35.6	35.6
	AIC and *F* test	66.0	65.4	65.4	66.0	67.2	64.2	66.8	66.6	65.2
Sensitive period	LARS	100	100	99.6	100	100	100	100	100	100
	F test	100	99.8	96.8	100	100	99.6	100	100	100
	AIC and *F* test	96.8	95.2	93.0	94.4	95.4	95.0	96.4	95.0	94.2
		n = 400			n = 1,000			n = 2,500		
$\sigma^2 = 9$ ($R^2 = 0.10$)	ρ	0	0.4	0.8	0	0.4	0.8	0	0.4	0.8
Critical period	LARS	97.2	97.0	92.2	98.2	97.4	97.2	96.6	96.8	97.8
	F test	67.0	64.6	58.4	68.4	66.4	64.8	66.4	66.4	64.0
	AIC and *F* test	83.2	81.0	72.8	81.4	81.8	80.8	83.6	81.4	79.0
Accumulation	LARS	93.6	92.8	90.2	94.4	92.4	90.2	91.6	93.0	91.2
	F test	40.8	37.4	38.4	40.0	38.4	33.8	38.4	35.6	35.6
	AIC and *F* test	66.0	65.4	65.4	66.0	67.2	64.2	66.8	66.6	65.2
Sensitive period	LARS	71.8	62.8	17.6	99.4	98.0	69.2	100	100	98.4
	F test	84.6	77.4	48.6	98.8	94.2	71.8	100	100	91.8
	AIC and *F* test	78.4	70.8	34.6	93.8	90.2	66.8	96.4	95.0	87.2
		n = 400			n = 1,000			n = 2,500		
$\sigma^2 = 24$ ($R^2 = 0.04$)	ρ	0	0.4	0.8	0	0.4	0.8	0	0.4	0.8
Critical period	LARS	95.2	93.0	79.8	98.2	97.2	92.8	96.6	96.8	97.8
	F test	62.0	59.4	48.4	68.0	65.4	56.6	66.4	66.4	61.2
	AIC and *F* test	79.4	76.0	62.6	81.4	81.6	74.4	83.6	81.4	77.2
Accumulation	LARS	88.8	90.2	71.8	94.4	92.4	88.0	91.6	93.0	91.2
	F test	40.8	37.4	37.8	40.0	38.4	33.8	38.4	35.6	35.6
	AIC and *F* test	65.2	65.2	58.6	66.0	67.2	64.0	66.8	66.6	65.2
Sensitive period	LARS	11.4	9.8	0.0	66.8	63.6	13.8	98.8	97.8	66.8
	F test	47.2	38.8	17.6	83.4	75.6	44.2	98.0	94.2	76.0
	AIC and *F* test	30.2	24.6	2.8	74.6	67.2	30.8	95.0	90.2	68.6

AIC indicates Akaike information criterion; LARS, least angle regression.

replication and to estimate causal effects. Advantages of the structured approach are that it requires the hypotheses to be carefully specified a priori, and can accommodate complex compound hypotheses that involve interactions. For example, Forsdahl[28] argued that a poor standard of living in early years followed by later life prosperity would increase the risk of arteriosclerotic disease. In this example, the highest risk would be seen in those who change exposure between earlier and later time periods. Other hypotheses, such as nonlinear accumulation, can also be investigated. The flexibility of our approach allows many hypotheses even if they are epidemiologically implausible: it is important to triangulate the statistical findings with knowledge about biological and social plausibility. If suggested hypotheses are thought to be "too complex," our procedure allows for retreat to simpler, more interpretable, hypotheses if necessary. In further investigation, only the identified hypothesis need be considered, allowing precise definition of the causal effect(s) of interest and reducing their number, leading to improved estimation by marginal structural or structural nested models.[7,29]

Our approach has the advantage that the selected hypothesis will always be easy to interpret, provided that interpretable hypotheses are proposed a priori. This is in contrast to methods that provide a plot and invite interpretation based on

TABLE 3. Percentage of 500 Simulations in Which a 95% Confidence Interval Contained the True Parameter Value, in Simulations with Three Binary Exposure Measurements, Calculated by Two Different Structured Approaches

		n = 400			n = 1,000			n = 2,500		
$\sigma^2 = 1$	ρ	0	0.4	0.8	0	0.4	0.8	0	0.4	0.8
Null	LARS	95.0	94.2	95.2	93.8	95.0	96.2	92.2	94.6	94.4
	AIC and/or *F* test	86.2	87.8	92.4	86.8	89.0	92.2	84.2	84.2	87.2
Critical period	LARS	95.6	96.2	96.2	96.0	96.2	95.6	94.8	94.4	94.0
	AIC and/or *F* test	95.6	96.2	96.2	96.0	96.2	95.6	94.8	94.4	94.0
Accumulation	LARS	95.2	96.0	97.0	97.4	97.6	97.0	94.4	93.8	94.0
	AIC and/or *F* test	95.2	96.0	97.0	97.4	97.6	97.0	94.4	93.8	94.0
		n = 400			n = 1,000			n = 2,500		
$\sigma^2 = 9$	ρ	0	0.4	0.8	0	0.4	0.8	0	0.4	0.8
Null	LARS	95.0	94.2	95.2	93.8	95.0	96.2	92.2	94.6	94.4
	AIC and/or *F* test	86.2	87.8	92.4	86.8	89.0	92.2	84.2	84.2	87.2
Critical period	LARS	95.6	95.6	89.2	96.0	96.2	94.4	94.8	94.4	94.0
	AIC and/or *F* test	95.6	95.8	89.2	96.0	96.2	94.4	94.8	94.4	94.0
Accumulation	LARS	95.4	95.8	92.2	97.4	97.6	96.0	94.4	93.8	94.0
	AIC and/or *F* test	95.0	95.8	92.2	97.4	97.6	96.0	94.4	93.8	94.0
		n = 400			n = 1,000			n = 2,500		
$\sigma^2 = 24$	ρ	0	0.4	0.8	0	0.4	0.8	0	0.4	0.8
Null	LARS	95.0	94.2	95.2	93.8	95.0	96.2	92.2	94.6	94.4
	AIC and/or *F* test	86.2	87.8	92.4	86.8	89.0	92.2	84.2	84.2	87.2
Critical period	LARS	93.2	90.2	83.5	96.0	95.2	90.0	94.8	94.4	93.0
	AIC and/or *F* test	93.2	90.8	83.8	96.0	95.2	90.0	94.8	94.4	93.0
Accumulation	LARS	94.8	90.2	70.4	97.4	97.4	89.2	94.4	93.8	93.6
	AIC and/or *F* test	94.8	89.2	70.4	97.4	97.4	89.0	94.4	93.8	93.6

AIC indicates Akaike information criterion; LARS, least angle regression.

curves on the graph.[3,4] The LARS algorithm has the advantage of choosing the most important hypothesis first: that corresponding to the greatest proportion of the variability in the outcome, while unimportant effects are deselected. Unlike the structured *F* test approach,[9] our approach can be employed without using *P* values. Using the covariance test for the lasso, we calculated confidence intervals with coverage unaffected by the fact that the selected hypothesis offers the best fit to the observed data. The covariance test for the lasso should be used in preference to an *F* test between simple and compound hypothesized models.

If the association between exposure and outcome is weak, with little variation in the outcome explained by the exposure, then the reliability of structured approaches in identifying the true model is diminished. It is therefore important first to examine the overall amount of variation explained by the exposure, perhaps using the R^2 value in a saturated model or extreme end of the elbow plot, and second to examine the correlation structure among the exposure measurements. If exposure measurements are highly correlated, it may be impossible to distinguish hypotheses without a large sample size or many repeated measures. Additional study is required to extend this approach to continuous exposures and categorical outcomes, consider measurement error in the exposure, and accommodate possible confounding.

Within the same overall structure, other variable selection methods might be used instead of the lasso. One possibility is the elastic net,[30] although this could select more variables than parameters in the saturated model, which would not be an advantage in understanding the hypothesized life course association, as less parsimonious models are less interpretable. The grouped lasso would allow all variables encoding a compound hypothesis to be selected at the same time,[31] and prevent only one component variable of a compound hypothesis being identified on its own. However, if compound hypotheses are misspecified, for example not all of their components are associated with the outcome, then not grouping variables still allows the important components to be extracted and the hypothesis to be identified.

Our conclusion is that the LARS procedure, implementing the lasso, can select hypotheses from a prespecified set, identifying the optimal hypothesis that offers the greatest consistency with the data. Compound hypotheses can be built from simpler hypotheses in a straightforward way, using

a small number of variables. The selected hypothesis and its potential causality may then be more precisely defined and further studied.

ACKNOWLEDGMENTS

The authors are grateful to the MRC Unit for Lifelong Heath and Ageing at University College London for providing the data used as an example.

REFERENCES

1. Barker DJ, Osmond C, Forsén TJ, Kajantie E, Eriksson JG. Trajectories of growth among children who have coronary events as adults. *N Engl J Med*. 2005;353:1802–1809.
2. Ben-Shlomo Y, Kuh D. A life course approach to chronic disease epidemiology: conceptual models, empirical challenges and interdisciplinary perspectives. *Int J Epidemiol*. 2002;31:285–293.
3. De Stavola BL, Nitsch D, dos Santos Silva I, et al. Statistical issues in life course epidemiology. *Am J Epidemiol*. 2006;163:84–96.
4. Tilling K, Howe LD, Ben-Shlomo Y. Commentary: methods for analysing life course influences on health–untangling complex exposures. *Int J Epidemiol*. 2011;40:250–252.
5. Tu YK, Tilling K, Sterne JA, Gilthorpe MS. A critical evaluation of statistical approaches to examining the role of growth trajectories in the developmental origins of health and disease. *Int J Epidemiol*. 2013;42:1327–1339.
6. Liu S, Jones RN, Glymour MM. Implications of lifecourse epidemiology for research on determinants of adult disease. *Public Health Rev*. 2010;32:489–511.
7. Petersen ML, van der Laan MJ. Causal models and learning from data: integrating causal modeling and statistical estimation. *Epidemiology*. 2014;25:418–426.
8. Pollitt RA, Rose KM, Kaufman JS. Evaluating the evidence for models of life course socioeconomic factors and cardiovascular outcomes: a systematic review. *BMC Public Health*. 2005;5:7.
9. Mishra G, Nitsch D, Black S, De Stavola B, Kuh D, Hardy R. A structured approach to modelling the effects of binary exposure variables over the life course. *Int J Epidemiol*. 2009;38:528–537.
10. Gustafsson PE, Janlert U, Theorell T, Westerlund H, Hammarström A. Fetal and life course origins of serum lipids in mid-adulthood: results from a prospective cohort study. *BMC Public Health*. 2010;10:484.
11. Birnie K, Martin RM, Gallacher J, et al. Socio-economic disadvantage from childhood to adulthood and locomotor function in old age: a life-course analysis of the Boyd Orr and Caerphilly prospective studies. *J Epidemiol Community Health*. 2011;65:1014–1023.
12. Murray ET, Mishra GD, Kuh D, Guralnik J, Black S, Hardy R. Life course models of socioeconomic position and cardiovascular risk factors: 1946 birth cohort. *Ann Epidemiol*. 2011;21:589–597.
13. Park MH, Sovio U, Viner RM, Hardy RJ, Kinra S. Overweight in childhood, adolescence and adulthood and cardiovascular risk in later life: pooled analysis of three british birth cohorts. *PLoS One*. 2013;8:e70684.
14. Evans J, Melotti R, Heron J, et al. The timing of maternal depressive symptoms and child cognitive development: a longitudinal study. *J Child Psychol Psychiatry*. 2012;53:632–640.
15. Kakinami L, Séguin L, Lambert M, Gauvin L, Nikiema B, Paradis G. Comparison of three lifecourse models of poverty in predicting cardiovascular disease risk in youth. *Ann Epidemiol*. 2013;23:485–491.
16. Richmond RC, Simpkin AJ, Woodward G, et al. Prenatal exposure to maternal smoking and offspring DNA methylation across the lifecourse: findings from the Avon Longitudinal Study of Parents and Children (ALSPAC). *Hum Mol Genet*. 2015;24:2201–2217.
17. Pearson RM, Melotti R, Heron J, et al. Disruption to the development of maternal responsiveness? The impact of prenatal depression on mother-infant interactions. *Infant Behav Dev*. 2012;35:613–626.
18. Wills AK, Black S, Cooper R, et al. Life course body mass index and risk of knee osteoarthritis at the age of 53 years: evidence from the 1946 British birth cohort study. *Ann Rheum Dis*. 2012;71:655–660.
19. Delgado-Angula EK, Bernabé E. Comparing lifecourse models of social class and adult oral health using the 1958 National Child Development Study. *Community Dent Hlth*. 2015;32:20–25.
20. Akaike H. A new look at the statistical model identification. *IEEE Transacton Automat Control*. 1974;19:716–723.
21. Mishra GD, Chiesa F, Goodman A, De Stavola B, Koupil I. Socio-economic position over the life course and all-cause, and circulatory diseases mortality at age 50-87 years: results from a Swedish birth cohort. *Eur J Epidemiol*. 2013;28:139–147.
22. Preston SH, Mehta NK, Stokes A. Modeling obesity histories in cohort analyses of health and mortality. *Epidemiology*. 2013;24:158–166.
23. Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc Ser B*. 1996;58:267–288.
24. Efron B, Hastie T, Johnstone I, Tibshirani R. Least angle regression. *Ann Stat*. 2004;32:407–499.
25. Radchenko P, James GM. Improved variable selection with forward-lasso adaptive shrinkage. *Ann Appl Stat*. 2011;5:427–448.
26. Davies PL, Kovac A. Local extremes, runs, strings and multiresolution. *Ann Stat*. 2001;29:1–65.
27. Lockhart R, Taylor J, Tibshirani R, Tibshirani R. A significance test for the lasso. *Ann Stat*. 2014;42:413–468.
28. Forsdahl A. Are poor living conditions in childhood and adolescence an important risk factor for arteriosclerotic heart disease? *Br J Prev Soc Med*. 1977;31:91–95.
29. Robins JM, Hernán MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology*. 2000;11:550–560.
30. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc Ser B*. 2005;67:301–320.
31. Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. *J R Stat Soc Ser B*. 2006;68:49–67.