American Journal of Epidemiology

Vol. 190, No. 6
DOI: 10.1093/aje/kwaa246
Advance Access publication:
March 26, 2021

Practice of Epidemiology

A Structured Approach to Evaluating Life-Course Hypotheses: Moving Beyond Analyses of Exposed Versus Unexposed in the -Omics Context

Yiwen Zhu, Andrew J. Simpkin, Matthew J. Suderman, Alexandre A. Lussier, Esther Walton, Erin C. Dunn*, and Andrew D. A. C. Smith

* Correspondence to Dr. Erin C. Dunn, Psychiatric and Neurodevelopmental Genetics Unit, Center for Genomic Medicine, Massachusetts General Hospital, 185 Cambridge Street, Simches Research Building, 6th Floor, Boston, MA 02114 (e-mail: edunn2@mgh.harvard.edu).

Initially submitted October 16, 2019; accepted for publication October 28, 2020.

The structured life-course modeling approach (SLCMA) is a theory-driven analytical method that empirically compares multiple prespecified life-course hypotheses characterizing time-dependent exposure-outcome relationships to determine which theory best fits the observed data. In this study, we performed simulations and empirical analyses to evaluate the performance of the SLCMA when applied to genomewide DNA methylation (DNAm). Using simulations ($n = 700$), we compared 5 statistical inference tests used with SLCMA, assessing the familywise error rate, statistical power, and confidence interval coverage to determine whether inference based on these tests was valid in the presence of substantial multiple testing and small effects—2 hallmark challenges of inference from -omics data. In the empirical analyses ($n = 703$), we evaluated the time-dependent relationship between childhood abuse and genomewide DNAm. In simulations, selective inference and the max-|*t*|-test performed best: Both controlled the familywise error rate and yielded moderate statistical power. Empirical analyses using SLCMA revealed time-dependent effects of childhood abuse on DNAm. Our findings show that SLCMA, applied and interpreted appropriately, can be used in high-throughput settings to examine time-dependent effects underlying exposure-outcome relationships over the life course. We provide recommendations for applying the SLCMA in -omics settings and encourage researchers to move beyond analyses of exposed versus unexposed individuals.

Avon Longitudinal Study of Parents and Children; DNA methylation; life course; -omics; postselection inference; structured approach

Abbreviations: ALSPAC, Avon Longitudinal Study of Parents and Children; ARIES, Accessible Resource for Integrated Epigenomic Studies; CpG, cytosine-phosphate-guanine; DNAm, DNA methylation; FWER, familywise error rate; LASSO, least absolute shrinkage and selection operator; SLCMA, structured life-course modeling approach.

Epidemiologists have long been interested in whether and how exposures incurred over the life course affect later health outcomes. Guided by theories developed in life-course epidemiology (Table 1), researchers are moving beyond simple comparisons of the presence versus absence of exposure to characterize time-dependent exposure-outcome relationships (1). Prior work in life-course epidemiology has conceptualized timing effects in numerous ways, examining the roles of the developmental timing of exposure (the sensitive-period hypothesis), the number of exposure occasions across time (the accumulation-of-risk hypothesis), proximity in time to exposure (the recency hypothesis), and change in exposure status across time (the mobility hypothesis). Researchers have adopted this life-course perspective, uncovering mechanistic insights that have advanced many subfields of public health and medicine (2–6). Because different life-course hypotheses correspond to distinct theories of disease etiology, efforts to formally compare competing hypotheses and identify those best supported by empirical data are needed to guide prevention and intervention planning.

To address the need for systematic comparisons of life-course theories, Mishra et al. (7) introduced the structured life-course modeling approach (SLCMA). The SLCMA

Table 1. Commonly Tested Life-Course Theories Characterizing Time-Dependent Relationships Between Exposures and Health Outcomes

Hypothesis	Life-Course Theory	Definition	Encoding[a]	Example[b]
Sensitive period	The developmental timing of exposure X has the strongest effect on the outcome at a specific time point due to heightened levels of plasticity or reprogramming.	Exposure at a particular time point j (X_j) is associated with the outcome.	X_j	$\text{Abuse}_{\text{period 1}}$ (X_1) = exposed (1) vs. unexposed (0) in time period 1
Accumulation	Every additional time point of exposure affects the outcome in a dose-response manner, independent of the exposure timing.	The accumulated sum of the number of exposure occasions (A) is linearly associated with the outcome.	$A = X_1 + \cdots + X_m$	$\text{Abuse}_{\text{accumulation}}$ (A) = number of time periods exposed to abuse (range, 0–6)
Recency	More proximal exposures (those that happen closer in time to the measurement of the outcome) are more strongly linked to the outcome than are more distal exposures.	The *weighted* sum (R) of the number of exposure occasions is linearly associated with the outcome such that the weight of each exposure is proportional to the age at the time of measurement.	$R = X_1T_1 + \cdots + X_mT_m$	$\text{Abuse}_{\text{recency}}$ (R) = $\text{abuse}_{\text{period 1}}$ exposed (1) vs. unexposed (0) × ($\text{age}_{\text{period 1}}$) + … + $\text{abuse}_{\text{period 6}}$ exposed (1) vs. unexposed (0) × ($\text{age}_{\text{period 6}}$)
Mobility	The change in exposure status between 2 time periods, rather than the absolute state at each individual time point, affects the outcome.	The unidirectional change (M_{jk}^{+} or M_{jk}^{-}) between 2 measurement occasions (from jth to kth) is associated with the outcome.	Positive change: $M_{jk}^{+} = (1 - X_j)X_k$ Negative change: $M_{jk}^{-} = X_j(1 - X_k)$	$\text{Abuse}_{\text{mobility}^{+}, \text{ period 1–2}}$ ($M_{1,2}^{+}$) = (1 − exposed (1) in time period 1) × exposed (1) in time period 2

[a] Notation is based on the description of hypotheses by Smith et al. (9). Let $X_1, \ldots, X_m$ be a set of m repeated binary measures of exposure (0 = unexposed; 1 = exposed) and $T_1 \ldots T_m$ the corresponding age at the time of measurement. X_j represents the measurement taken on the jth measurement occasion.

[b] This column shows examples of how the life-course theories, which were tested in empirical analyses of the epigenomewide structured life-course modeling approach to examine exposure to physical or sexual abuse in childhood, could be encoded. Notably, the accumulation models can also be parameterized differently, such as with nonlinear effects ("U-shaped" or "J-shaped" relationships). However, for simplicity, we provide the simplest definition of accumulation, which is also the most frequently tested.

allows researchers to compare a set of a priori–specified life-course theories and use goodness-of-fit criteria to determine which theory is best supported by empirical data. Smith et al. (8) later extended this approach with an alternative statistical model selection strategy that uses least-angle regression, accommodates both binary and continuous exposures (9, 10), and improves the accuracy of selecting the correct hypothesis. More recently, Madathil et al. (11) proposed a Bayesian approach to life-course modeling that does not perform variable selection but rather estimates the posterior probability corresponding to each theoretical hypothesis while assessing the relative importance of a series of life-course theories. Since its inception, the SLCMA has been applied in a wide range of non-omics epidemiologic studies, including those examining the time-dependent impacts of childhood trauma, physical activity, or socioeconomic position on psychological, metabolic, and disease outcomes (12–18). Compared with other approaches that consider alternative classifications of the exposure, the SLCMA is better positioned to compare competing life-course hypotheses simultaneously. By requiring that life-course hypotheses be specified a priori on the basis of theory, it prevents post-hoc hypothesis-generation following exploratory analyses. Moreover, its model selection feature allows a structured assessment of hypotheses without requiring a saturated model.

The growing availability of high-dimensional biological and phenotypic data from longitudinal cohort studies has created new opportunities to assess time-varying exposures in epigenomics, transcriptomics, metabolomics, and other -omics settings (19–21). While large cross-sectional -omics studies have identified *associations* between biological differences and various traits (22), applications of the SLCMA to longitudinal data and high-dimensional outcomes allow

researchers to answer more complex questions about disease *mechanisms*. For example, Dunn et al. (23) applied the SLCMA in a longitudinal birth cohort study to model timing effects of childhood adversity on DNA methylation (DNAm), which is a widely studied epigenetic mechanism that could give rise to altered gene expression and phenotypic changes. Using the SLCMA, they found that differences in DNAm were largely explained by age at exposure, with the first 3 years of life appearing to be a sensitive period associated with more DNAm differences. Their results also showed that the SLCMA could identify associations not identified by an epigenomewide association study of persons exposed to childhood adversity versus those unexposed (23), underscoring the importance of alternative exposure classifications.

In this study, we aimed to extend these findings with methodological contributions that outline the structured life-course modeling framework and its application in -omics settings. As discussed in Dunn et al. (23), application of the SLCMA to -omics data presents unique challenges not yet systematically investigated. First, it remains unknown whether theoretical properties of statistical inference, such as type I error (i.e., the familywise error rate (FWER) in the presence of multiple testing) or confidence interval coverage, are valid in -omics data. Second, it is unclear whether the SLCMA is sufficiently powered to detect the small effects commonly found in -omics settings. Third, questions exist on how to balance decision-making regarding research evidence, because -omics studies often rely on P values and accurate statistical inference has become increasingly important. Moreover, epidemiologists and other researchers increasingly prioritize other statistical evidence, such as effect sizes and confidence intervals (24, 25). We therefore performed simulations and empirical analyses to assess the performance of the SLCMA when applied to -omics data. We illustrate how the SLCMA can be applied to evaluate the time-dependent role of childhood abuse in genomewide DNAm.

METHODS

Overview of the SLCMA

The SLCMA has been described in detail elsewhere (7, 9, 10). In brief, the SLCMA is a 2-stage method that compares a set of life-course hypotheses describing the relationship between exposures assessed over time and some outcome of interest. In the first stage of the SLCMA, each life-course hypothesis is encoded into a predictor or set of predictor variables. Table 1 shows examples of predictors representing commonly studied life-course hypotheses. A variable selection procedure is then used to select the subset of predictors that explains the greatest proportion of outcome variation. While it is possible for multiple predictors to be selected, the high dimensionality of the -omics setting makes consideration of simple life-course hypotheses (meaning those in which the exposure-outcome association is represented by a single predictor) more feasible. Therefore, in this study, we focused on statistical inference regarding the single predictor explaining the greatest variation in the outcome.

In the second stage of the SLCMA, postselection inference is performed to obtain point estimates and confidence intervals for the model identified in the first stage. Postselection inference methods are used to derive unbiased test statistics because they account for the multiple testing that occurs when comparing multiple hypotheses (meaning the multiple testing occurring at the first stage, instead of the number of outcomes examined), as the SLCMA iteratively works to *select* the variable with the strongest association with the outcome. Four inference methods that account for this "selective nature" are 1) Bonferroni correction, 2) the max-|t|-test (26), 3) the covariance test (27, 28), and 4) selective inference (29, 30). These approaches are described in detail in Web Appendix 1 and Web Table 3 (available online at https://doi.org/10.1093/aje/kwaa246).

Simulation analyses

We conducted simulations to examine the performance of these 4 postselection inference methods as compared with a naive calculation (summarized in Table 2). To build these simulations in the context of real-world applications, we modeled the simulation strategy based on the genomewide SLCMA study performed by Dunn et al. (23). We evaluated each postselection inference method with respect to 3 statistical properties: the FWER (the probability of making 1 or more false discoveries out of multiple tests), statistical power (the probability of correctly selecting the predictor with a true association with the outcome), and confidence interval coverage (the probability that a 95% confidence interval contains the true effect estimate). Assessing these properties enabled us to determine whether inference based on these tests was valid in the presence of multiple testing and small effect sizes, which are 2 hallmarks of high-dimensional data. Mathematical definitions of the test statistics and the procedure for constructing confidence intervals, as well as example R code, are included in Web Appendices 1 and 2 and are available on GitHub (31). All postselection inference methods, including the naive calculations, involved multiple testing correction for the number of cytosine-phosphate-guanine (CpG) sites tested using a Bonferroni correction (i.e., the P value threshold was $P < 1 \times 10^{-7}$).

Setup of simulations

We considered 2 scenarios, which differed in terms of the simulated outcome. In both scenarios, we simulated exposure to childhood sexual or physical abuse based on empirical data collected during 1991–2000 in the Avon Longitudinal Study of Parents and Children (ALSPAC), a population-based study of an English birth cohort (32–34). Pregnant women with estimated delivery dates between April 1991 and December 1992 were invited to be part of ALSPAC. We analyzed data from an ALSPAC subsample, the Accessible Resource for Integrated Epigenomic Studies (ARIES). We set our sample size to 700 to be consistent with ARIES. Simulations were based on 485,000 tests corresponding to an analysis of Illumina Infinium HumanMethylation450K BeadChip data (Illumina, Inc., San Diego, California). In scenario 1, the outcome (i.e., DNAm) was simulated from

Table 2. Setup of a Simulation Study Assessing the Performance of 5 Statistical Inference Tests Used With the Structured Life-Course Modeling Approach[a]

Simulation Parameter		
Under the Null (FWER)	**Outcome**	**No. of Tests**
Normal outcomes	$y \sim \mathcal{N}(0, 1)$	485,000
Empirical outcomes	Resampled DNAm values	485,000
Under the Alternative (Power and CI Coverage)	**Outcome**	**Effect Size[b]**
Normal outcomes	Simulated normal variables associated with the first predictor (earliest sensitive period)	R^2: 0.01–0.1
Empirical outcomes	Simulated β variables associated with the first predictor (earliest sensitive period)	Δ_{DNAm}: 0.05–0.5

Abbreviations: CI, confidence interval; DNAm, DNA methylation; FWER, familywise error rate.

[a] The table shows 2 different approaches to simulations of life-course modeling in the -omics context under the null and alternative settings: To assess the FWER, we simulated the exposures and outcomes to have no association with each other (i.e., under the null hypothesis) and ran a single simulation of 485,000 tests to examine the distributions of observed *P* values as compared with the expected distribution. To assess statistical power and CI coverage under the alternative hypothesis, we ran 2,000 simulation experiments to allow the CI of the assessed metrics (i.e., power and CI coverage) to have a radius (i.e., margin of error) of 1%, setting α to 5%. The 2 metrics of effect sizes were different with normal versus empirical outcomes because of the difference in the underlying data-generating processes. The sample size was set to $n = 700$ in all simulations based on the sample size of the empirical study. For all simulation analyses, the predictors were simulated on the basis of exposure to childhood abuse from the Avon Longitudinal Study of Parents and Children (England, 1991–2000). The analyses included 7 variables encoding sensitive period, accumulation, and recency hypotheses.

[b] R^2: variance of the outcome explained by the selected predictor; Δ_{DNAm}: difference in average DNAm levels between exposed and unexposed individuals.

a normal distribution. In scenario 2, we resampled the outcomes under the null to more closely resemble "β" values, which represent the proportion of cells in which the cytosine at the locus is methylated and range from 0 to 1. To assess statistical power and confidence interval coverage, we simulated the outcome from a beta distribution, as proposed by Tsai and Bell (32). In both scenarios, the effect sizes were selected to illustrate a wide range of statistical power based on previous epigenomewide association studies examining different exposures (33, 34).

To assess model misspecification, we also conducted simulations in which 1) the outcome variable was correlated with a variable encoding an alternative hypothesis (ever exposed vs. never exposed) not included in the prespecified set of hypotheses tested and 2) the outcome variable was correlated with 2 predictors (a compound life-course hypothesis). We also varied the sample size to investigate its effect on statistical power.

Full details of the simulations are provided in Web Appendix 1.

Measurement of power and confidence interval coverage

Conceptually, bias might arise from the SLCMA analysis in 2 ways. First, at the first stage, the model most supported by the sample data may not be the model most supported in the population. At the second stage, even if the model has been correctly selected, inference based on that model may be biased. In our simulations, we considered both uncertainties residing in model selection and inference: Power was calculated as the percentage of times that the first (variable selection) stage correctly selected the model and the second (inference) stage identified it as a below-threshold hit. Similarly, confidence interval coverage was calculated as the percentage of times that the first stage correctly selected the model and the confidence interval contained the true value. Alternatively, if the first stage selected the wrong model but the confidence interval contained 0, we considered that the true effect (since there should be no effect) was captured by the confidence interval.

Empirical analyses

To illustrate how the SLCMA and the different corresponding postselection inference methods work in practice, we reanalyzed the data used by Dunn et al. (23). Briefly, we compared the effects of sensitive period, accumulation, and recency hypotheses for the associations between exposure to sexual or physical abuse and genomewide DNAm at age 7 years among ALSPAC participants ($n = 703$). Sample characteristics and adversity measures are described in Web

Table 3. Statistical Properties of Postselection Inference Methods (Main Findings) in Simulated Epigenomewide Analyses of the Time-Dependent Relationship Between Childhood Abuse and DNA Methylation (n = 700)[a]

Method	FWER (Figures 1 and 2)	Statistical Power (Figure 3)	CI Coverage (Figure 4)	Software Availability	Computation Time for an Epigenomewide Analysis[b]
Naive calculation	Inflated *P* values and FWER	Biased due to inflated FWER	Lower-than-expected coverage when effect size is small (9)	Widely available	Fast (24 minutes)
Bonferroni correction	Controlled at any level	Comparable	Overly conservative (i.e., above expected coverage)	Widely available	Fast (24 minutes)
Max-|*t*|-test	Controlled at any level	Comparable	Lower-than-expected coverage when effect size is small	R code provided in Web Appendix 2	Slow (11 hours and 51 minutes)
Covariance test	Inflated *P* values and FWER	Biased due to inflated FWER	Expected coverage (9); interval not necessarily contiguous	R package archived (28)	Moderate (1 hour and 19 minutes)
Selective inference	Controlled at any level	Comparable	Expected coverage	R package available (30); possible to implement generalized linear models as well	Slow (14 hours and 13 minutes)

Abbreviations: CI, confidence interval; FWER, familywise error rate.
[a] Simulations were based on data obtained from the Avon Longitudinal Study of Parents and Children (England, 1991–2000).
[b] Computation time was based on analyses running under R 3.4.0 (R Foundation for Statistical Computing, Vienna, Austria) using a high-performance computer cluster with 8 GB of random access memory and a maximum of 6 central processing unit cores allotted.

Appendix 3. Building from that study, which used only the covariance test, we additionally applied the other postselection inference methods summarized above.

The most widely used covariate adjustment strategy in the SLCMA is to regress the exposures on the covariates and enter the residuals into variable selection, which decreases the likelihood that observed associations are due to measured confounders. We also tested a new method for covariate adjustment that could be used alongside any postselection inference method. Based on the Frisch-Waugh-Lovell theorem, this method also regresses the outcome on covariates and enters the residuals into the model selection procedure (35–37). A thorough description of this method and the full list of covariates are available in Web Appendix 1. Notably, the SLCMA requires a common set of confounders to be prespecified for all hypotheses; thus, bias may arise from time-varying or hypothesis-dependent confounding.

RESULTS

Simulation analyses

Table 3 summarizes the main findings from the simulation analyses regarding the statistical properties and implementation of the assessed methods.

Familywise error rate

Because of the high computational burden of genomewide association studies, we illustrated FWER control of each inference test using a single simulation with m = 485,000 tests. As Figures 1 and 2 show, when compared against the expected *P*-value distribution under the null hypothesis, the *P* values obtained from naive calculations appeared to be too liberal in both scenarios, as suggested by the systematic upward departure from the diagonal line. *P* values from the covariance test were also smaller than expected across scenarios.

With normally distributed outcomes in scenario 1, the *P* values from the Bonferroni correction, the max-|*t*|-test, and the selective inference method followed the expected distributions closely (Figure 1). With empirical DNAm outcomes in scenario 2, *P* values from the 3 methods seemed conservative (Figure 2). Transforming the DNAm (β) values to *M* values did not affect the results (Web Figure 1). Together, these findings suggest that 3 methods adequately controlled the FWER: Bonferroni correction, the max-|*t*|-test, and the selective inference method. Web Appendix 1 and Web Table 4 show estimates of FWER obtained from repeated simulation experiments when the number of tests ranged from m = 1 to m = 1,000.

Statistical power and confidence interval coverage

We assessed the statistical power of the 3 methods that adequately controlled FWER. We did not evaluate the performance of the covariance test or naive calculation, as these methods would have their statistical power unfairly inflated by their tendency to fail to reject the null hypothesis.

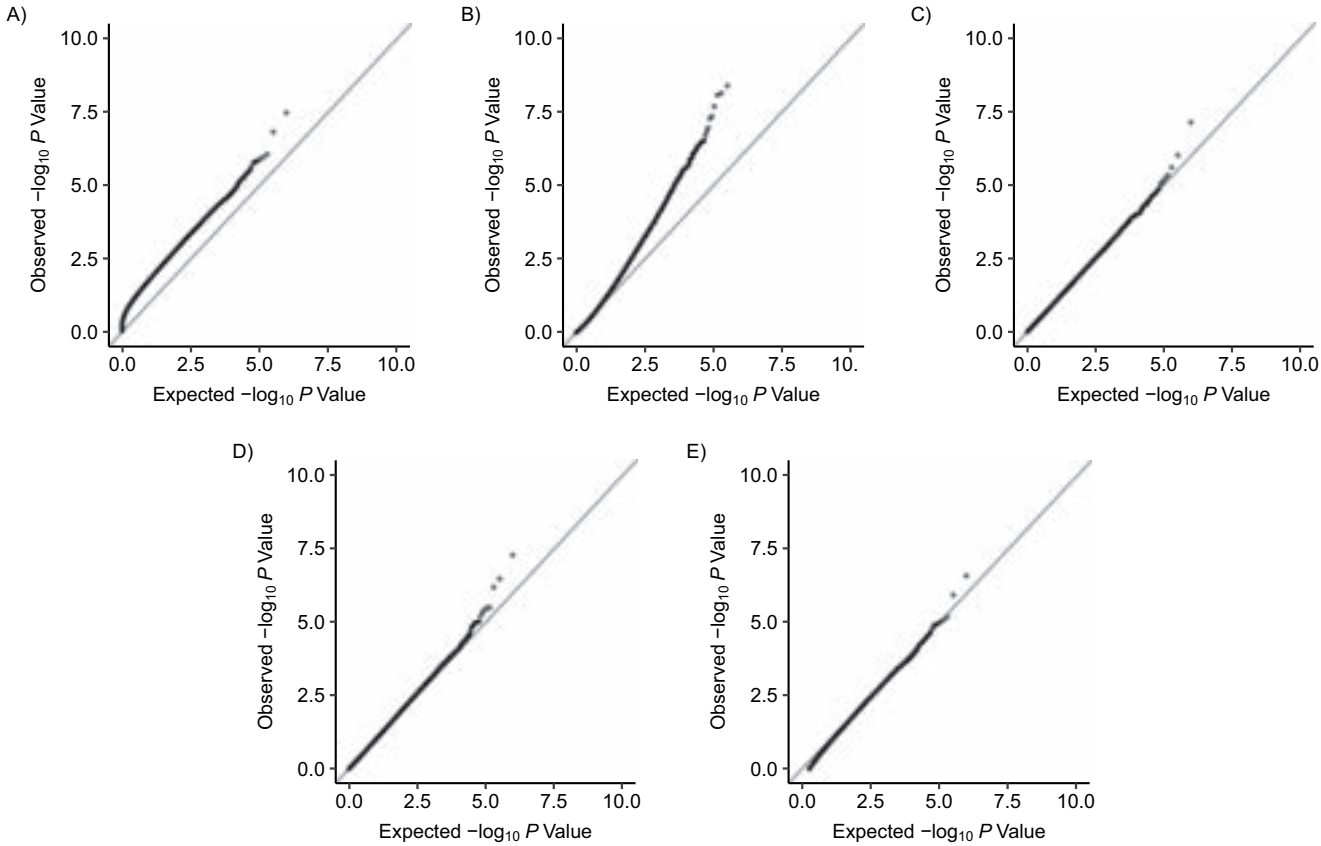


Figure 1. Q-Q plots comparing the expected *P* values with the observed *P* values simulated under the null for naive calculations and 4 postselection inference methods (n = 700) with normal outcomes, where the outcome variables were simulated to follow a normal distribution (scenario 1). A) Naive calculations; B) covariance test (27); C) selective inference (29); D) max-|*t*|-test (26); E) Bonferroni correction. Simulations were based on data obtained from the Avon Longitudinal Study of Parents and Children (England, 1991–2000).

Results suggested there was very little difference in statistical power between the 3 methods (Figure 3); they all had ideal statistical power (over 80%) when the effects were moderate to large ($R^2 > 0.06$ in scenario 1; $\Delta_{\text{DNAm}} > 0.25$ in scenario 2). With normal outcomes, selective inference achieved ideal confidence interval coverage (around 95%) across all effect sizes with sample size n = 700; the max-|*t*|-test had slightly lower coverage when the effect size was small ($R^2 < 0.03$). With outcomes simulated from beta distributions, the confidence interval coverage probabilities were below the desired level (95%) when the between-group difference (Δ_{DNAm}) was below 0.3, though they exceeded 95% as the effect size increased. Bonferroni-corrected confidence intervals were overly conservative across effect sizes and scenarios, as expected (Figure 4).

Robustness to model misspecification

If none of the predictors represent the true underlying life-course hypothesis, a misspecified model may be selected. In our simulations of this case, we found that the accumulation or recency model was often selected, because these models were highly correlated with the true predictor—ever being exposed versus never being exposed ($r_{\text{accumulation}}$ = 0.89, r_{recency} = 0.82). However, the power was reduced in comparison with a correctly specified model (Web Figure 2). If the true hypothesis is represented by 2 or more predictors (i.e., a compound hypothesis), the power to select 1 of these predictors may be diminished. In our simulations, the power to select 1 predictor was lower for selective inference (Figure 5). However, selective inference is the only method available for postselection inference on the second predictor that does not inflate the FWER. Statistical power increased with sample size for all methods considered (Web Figure 3).

Empirical analyses

Using the covariance test, Dunn et al. (23) identified 5 CpG sites in ALSPAC that showed differential methylation profiles at age 7 years following exposure to physical or sexual abuse in childhood; the "sensitive period" model was the selected life-course theory for these 5 sites. We performed the genomewide SLCMA analyses using 2 other postselection inference methods that showed no inflation in FWER and desired confidence interval coverage: the max-|*t*|-test and the selective inference method. Results are

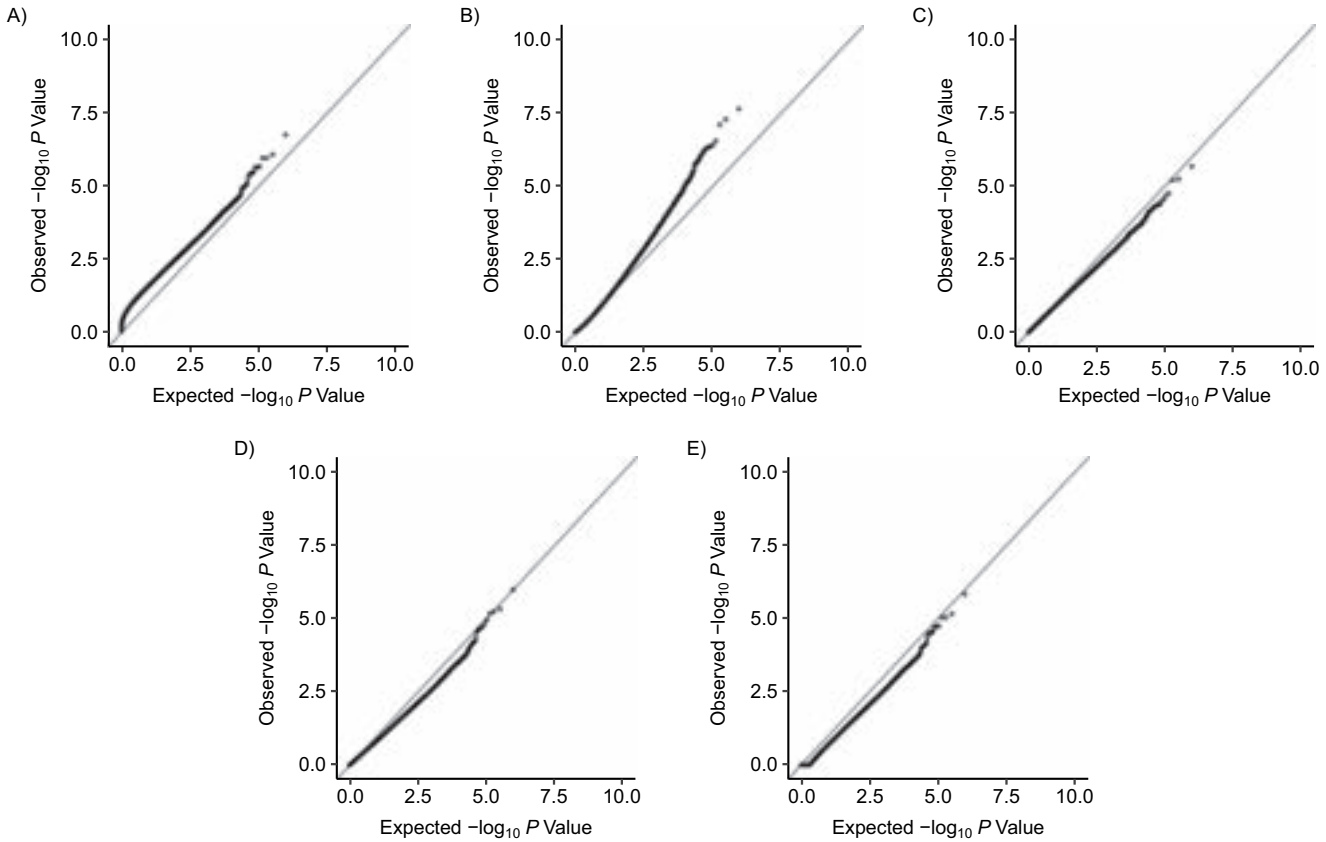


Figure 2. Q-Q plots comparing the expected *P* values with the observed *P* values simulated under the null for naive calculations and 4 postselection inference methods (n = 700) with empirical outcomes, where the outcome variables were resampled from observed DNA methylation values (scenario 2). A) Naive calculations; B) covariance test (27); C) selective inference (29); D) max-|*t*|-test (26); E) Bonferroni correction. Simulations were based on data obtained from the Avon Longitudinal Study of Parents and Children (England, 1991–2000).

shown in Web Table 1. While neither method identified any CpG site as significantly associated using a stringent Bonferroni-corrected *P* value threshold of $P < 1 \times 10^{-7}$, the CpG site with the smallest *P* value from the covariance test (cg06430102) remained the CpG with the smallest *P* value (out of the 485,000 CpG sites tested) for the 2 alternative methods (Web Table 1). The confidence intervals calculated on the basis of the covariance test, selective inference, and the max-|*t*|-test substantially overlapped (Figure 6; Web Table 1). On a genomewide level, concordance between the liberal covariance test and the recommended selective inference method was high, implying that both methods agreed on the loci that had the strongest associations with exposure (Web Table 2).

After we applied the Frisch-Waugh-Lovell theorem to additionally adjust for covariates, the *P* values decreased at all 5 loci (Web Figure 4), suggesting that the approach improved statistical power while retaining control for confounding (Web Figure 5).

DISCUSSION

As the availability of longitudinal biological and phenotypic data grows in the era of "big data," combining -omics technologies with rigorous epidemiologic methods can reveal critical new knowledge about biological mechanisms (38–40). Specifically, methods from life-course epidemiology can be translated to "harness the 'omics' revolution" (2, p. 984) and provide insights into how exposures become biologically embedded. We showed that, under a set of untestable assumptions, one such method—the SLCMA—can be used to directly compare life-course theories and can be scaled up to answer nuanced questions about time-dependent exposure-omics relationships. For example, if an early childhood sensitive-period hypothesis was selected for a locus known to be implicated in circadian rhythms, this finding could point to ways in which the biological clock is influenced by exposures during periods of heightened plasticity. If the accumulation hypothesis was selected for most of the loci implicated in inflammation, this finding could suggest dose-response relationships between the exposure and inflammatory responses.

Importantly, not all SLCMA methods for statistical inference are suitable in high-throughput applications. Our findings recommend the selective inference method and the max-|*t*|-test for postselection inference in -omics applications. Our simulations also showed that statistical power to detect effects depended on effect size but not necessarily

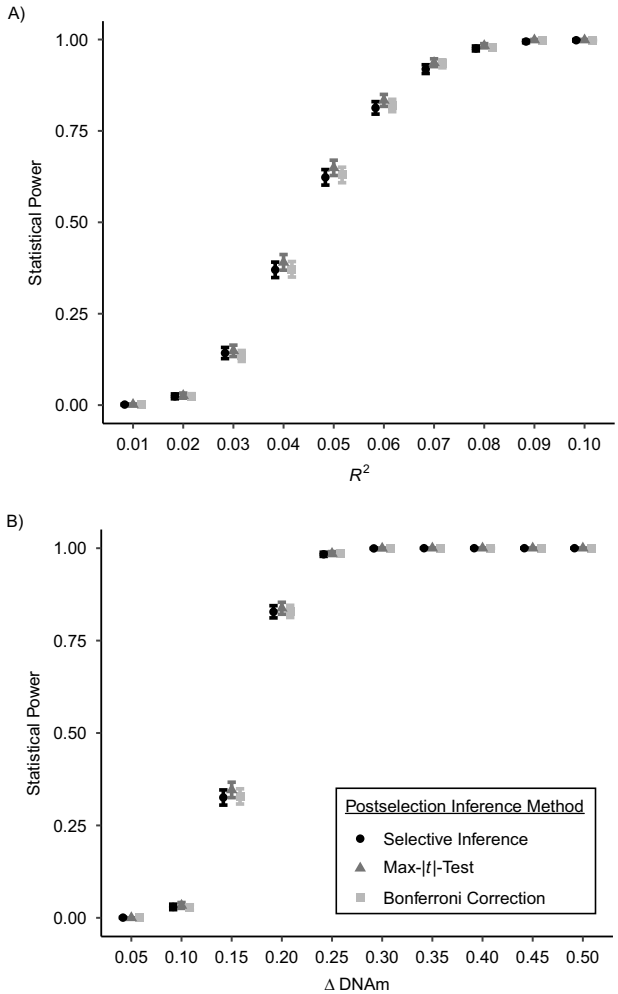


Figure 3. Estimated statistical power in simulated epigenomewide analyses of the time-dependent relationship between childhood abuse and DNA methylation (n = 700), with varying effect sizes. A) Normal outcomes; B) beta-distributed outcomes. Technical details about selective inference (29) and the max-|t|-test (26) are provided in Web Appendix 1. Simulations were based on data obtained from the Avon Longitudinal Study of Parents and Children (England, 1991–2000). Bars, 95% confidence intervals. DNAm, DNA methylation.

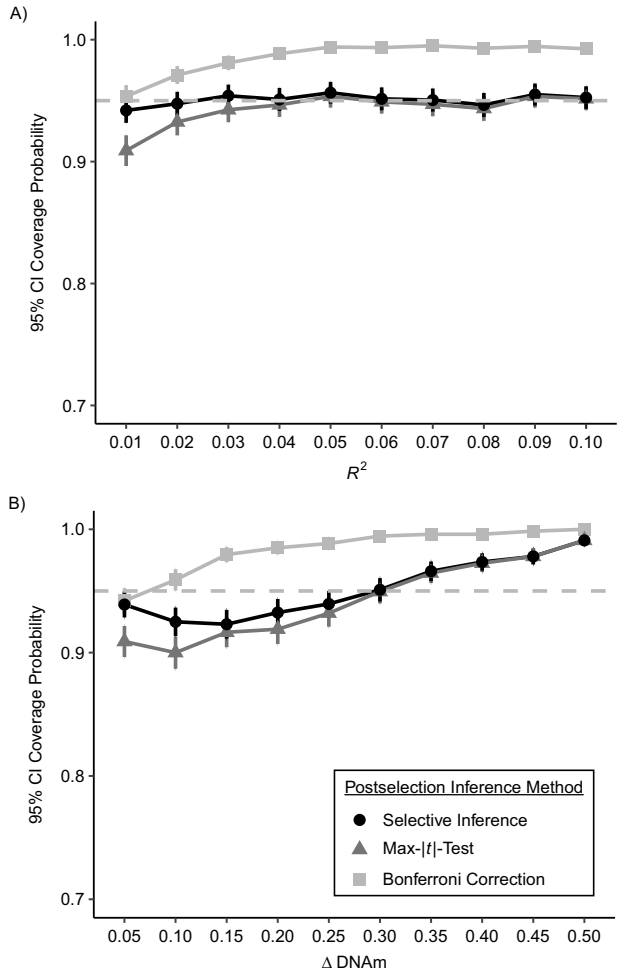


Figure 4. Estimated confidence interval (CI) coverage probability in simulated epigenomewide analyses of the time-dependent relationship between childhood abuse and DNA methylation (n = 700), with varying effect sizes. The gray dashed line corresponds to the prespecified coverage probability (95%). A) Normal outcomes; B) beta-distributed outcomes. Technical details about selective inference (29) and the max-|t|-test (26) are provided in Web Appendix 1. Simulations were based on data obtained from the Avon Longitudinal Study of Parents and Children (England, 1991–2000). Bars, 95% CIs. DNAm, DNA methylation.

on the postselection inference method used. When deciding between these 2 inference methods, researchers will need to consider several factors, including analytical goals and study-specific contexts, as both methods have strengths and limitations in these areas (Web Appendix 1). The simulation analyses highlight the value of using simulations in scientific research (41, 42), especially when theoretical assumptions may be violated in a new-application setting.

The empirical example presented in this paper extended the analyses performed by Dunn et al. (23), using 1 of the exposures and the same DNAm data (Web Appendix 3). However, these analyses differed by considering 2 alternative postselection inference methods (selective inference and the max-|t|-test) in the simulations. Comparing the covariance test with these 2 methods, we showed that statistical significance based on P values may differ across methods. The main reason for the discordance between the max-|t|-test and the 2 least absolute shrinkage and selection operator (LASSO)-based tests is that the max-|t|-test considers only the first predictor selected, whereas the selected inference is based on LASSO models that also consider subsequent predictors. Researchers should assess P values in parallel with effect estimates and confidence intervals, as decision rules of significance based on P values of 1 method may be biased due to inflation or overcorrection. Triangulating evidence from multiple sources and methods may suggest directions for future replication (43). For example, a CpG that was identified as the top site by multiple methods and

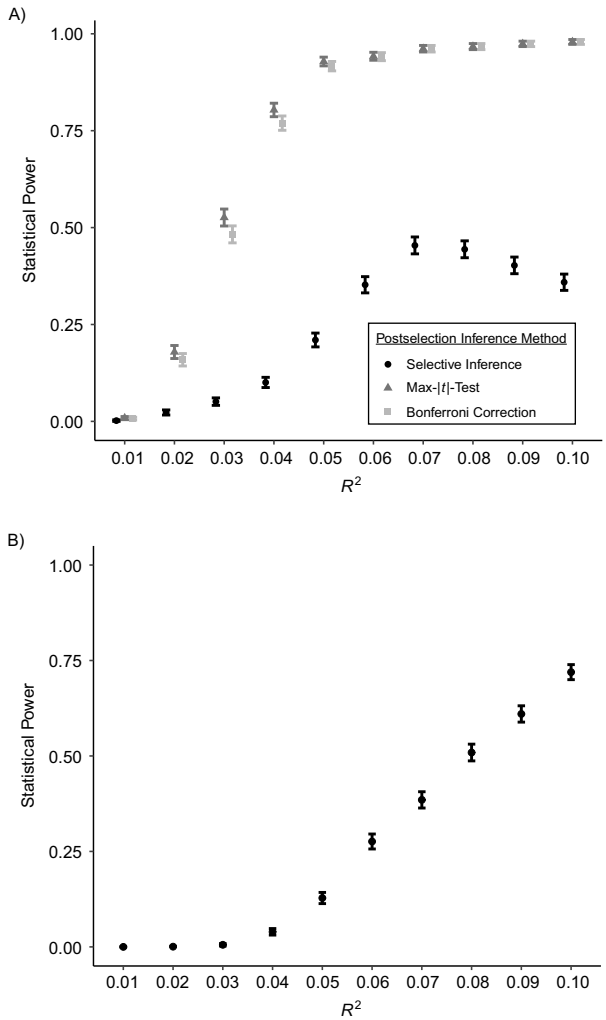


Figure 5. Estimated statistical power in simulated epigenomewide analyses of the time-dependent relationship between childhood abuse and DNA methylation (n = 700), with varying effect sizes, when the true causal relationship was represented by 2 hypotheses working in combination. A) Statistical power of selection of the first hypothesis (n = 700), when the true hypothesis is a compound hypothesis; B) statistical power of selection of the second hypothesis (n = 700), when the true hypothesis is a compound hypothesis. Technical details about selective inference (29) and the max-|t|-test (26) are provided in Web Appendix 1. Simulations were based on data obtained from the Avon Longitudinal Study of Parents and Children (England, 1991–2000). Bars, 95% confidence intervals.

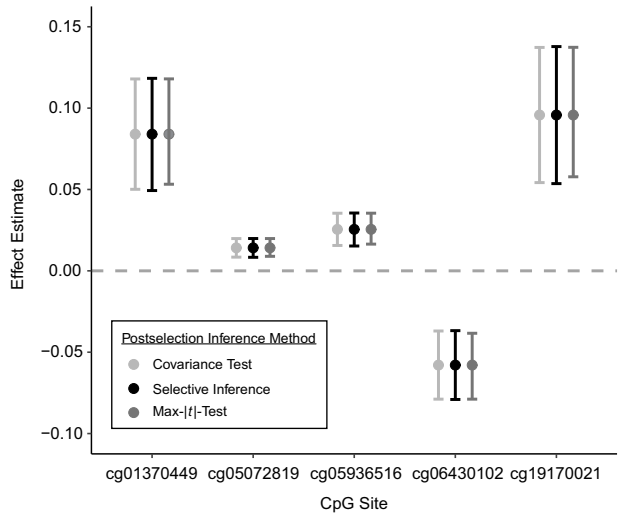


Figure 6. Overlap between confidence intervals based on the covariance test, selective inference, and the max-|t|-test in the empirical example examining the time-dependent relationship between childhood abuse and genomewide DNA methylation, showing the top 5 loci. Technical details about the covariance test (27), selective inference (29), and the max-|t|-test (26) are provided in Web Appendix 1. Simulations were based on data obtained from the Avon Longitudinal Study of Parents and Children (England, 1991–2000). Bars, 95% confidence intervals. CpG, cytosine-phosphate-guanine.

showed substantial changes in methylation levels between exposed and unexposed individuals may be more likely to capture effects of the exposure and worth pursuing in experimental validation.

Like any statistical method aspiring to address causal questions, the SLCMA relies on the usual assumptions that the model is correctly specified and that there is no unmeasured confounding (44). In simulations, we showed that when the model is misspecified, the SLCMA will identify hypothesized models with predictors that are correlated with the true model's predictors, but with reduced power. Therefore, SLCMA users must recognize that the selected hypothesis simply explains the most variation out of the (combinations of) candidate hypotheses considered, and there may be another (or nontested) theoretical model that explains more variation. Thus, careful formulation of the hypotheses is critical to capture the most plausible causal relationship based on prior literature or reasoning; consideration of alternative hypotheses (beyond those already selected) is also needed as research evidence grows. We would also emphasize that the selection of life-course models is based both on proper specification of the relevant hypothesis and on the set of candidate hypotheses included. For example, in our set of candidate hypotheses, we considered 1 sensitive period per time point when the exposure was measured. This approach may be inappropriate when the measurements are assessed close together in time: For example, for some exposure-outcome pairs, we might not claim to distinguish a sensitive period at 1.5 years from one at 2.5 years. In such cases, we recommend possibly condensing measurements into longer sensitive periods, taking the average exposure over all measurements in a time period defined by prior literature or reasoning. Such an approach increases the statistical power of variable selection procedures by reducing the number of and correlation between predictors.

The SLCMA has some limitations beyond the usual assumptions: In the current study, we assumed that the true hypothesis was represented by a single predictor (i.e., a simple hypothesis). Identifying more complicated exposure-outcome relationships in -omics settings may be of interest

but will require large sample sizes to achieve sufficient power. Moreover, the SLCMA currently does not accommodate time-varying confounding. It also does not allow for a different set of confounders for each hypothesis. In the empirical analyses, we tried to include a comprehensive set of baseline covariates based on prior literature that may be related to both childhood abuse and epigenetic changes. In light of these issues, the current results should be interpreted as suggestive evidence of loci that warrant future examination and replication in other data sets. Efforts to incorporate time-varying confounding into the SLCMA, such as marginal structural models (45, 46), are also needed.

Several other limitations of the current study are notable. First, although we varied the effect size and compared normal distributions of the outcome variable with empirical distributions of the outcome variable, we did not vary the distribution or correlation of the exposures, because of the number of possible combinations of these parameters. Thus, we encourage researchers to perform their own simulations to better understand the statistical properties of the SLCMA in their specific research context. Second, we restricted our analyses to linear-regression–based model selection; a brief discussion on the possibility of implementing postselection inference methods for generalized linear models is included in Web Appendix 1. Third, as suggested by the simulations, a typical longitudinal epigenetic study with a sample size under 1,000 is probably underpowered to detect small effects. In particular, when studying psychosocial exposures such as childhood abuse, we would not expect the exposure to have a large effect on DNAm at a single locus. For instance, power would likely be under 50% and confidence interval coverage may be lower than 95% when the outcome variation explained is below 5%, which has been common in previous epigenomewide association studies. One approach for improving statistical power is to combine data or summary results from multiple samples and perform a mega-/meta-analysis; development of methods for meta-analyzing results from SLCMA analyses is an important goal of future work. Another approach is to use the Frisch-Waugh-Lovell theorem for covariate adjustment, which leads to improvement in power, as we have shown in this paper. Fourth, the current SLCMA framework in the -omics setting does not restrict or penalize any loci based on their biological significance. One promising direction of future research is to leverage functional or regulatory information about the genomic regions under consideration (47, 48), especially when developmental stage-specific knowledge is available, in order to improve power and gain biological insights.

In conclusion, the SLCMA is a useful approach that brings the life-course perspective into the -omics context. Compared with an analysis that only categorizes exposure status as exposed versus unexposed, the SLCMA not only offers additional mechanistic insights about exposure mechanisms but also increases statistical power when the true underlying exposure-outcome relationship is more nuanced (23). As a field, epidemiology should move beyond analyses of the presence versus absence of exposure and make full use of repeatedly measured phenotypic and -omics data to generate knowledge that improves human health over the life course.

ACKNOWLEDGMENTS

Author affiliations: Center for Genomic Medicine, Massachusetts General Hospital, Boston, Massachusetts, United States (Yiwen Zhu, Alexandre A. Lussier, Erin C. Dunn); School of Mathematics, Statistics and Applied Mathematics, National University of Ireland, Galway, Ireland (Andrew J. Simpkin); MRC Integrative Epidemiology Unit, School of Social and Community Medicine, University of Bristol, Bristol, United Kingdom (Matthew J. Suderman, Esther Walton); Department of Psychiatry, Harvard Medical School, Boston, Massachusetts, United States (Alexandre A. Lussier, Erin C. Dunn); Department of Psychology, Faculty of Humanities and Social Sciences, University of Bath, Bath, United Kingdom (Esther Walton); Stanley Center for Psychiatric Research, Eli and Edythe L. Broad Institute of MIT and Harvard, Cambridge, Massachusetts, United States (Erin C. Dunn); Henry and Alison McCance Center for Brain Health, Massachusetts General Hospital, Boston, Massachusetts, United States (Erin C. Dunn); and Applied Statistics Group, University of the West of England, Bristol, United Kingdom (Andrew D. A. C. Smith).

Both senior authors (E.C.D. and A.D.A.C.S.) contributed equally to this work.

This work was supported by the National Institute of Mental Health, US National Institutes of Health (grant R01MH113930 awarded to E.C.D.). E.W. was supported by the Cohort and Longitudinal Studies Enhancement Resources (CLOSER) Consortium, which is funded by the Economic and Social Research Council and the Medical Research Council (MRC) (grant ES/K000357/1). The MRC and the Wellcome Trust (grant 102215/2/13/2) and the University of Bristol provide core support for the Avon Longitudinal Study of Parents and Children (ALSPAC). Creation of the Accessible Resource for Integrated Epigenomic Studies (ARIES), comprising a subsample of ALSPAC participants, was funded by the Biotechnology and Biological Sciences Research Council (grants BBI025751/1 and BB/I025263/1). Supplementary funding for generation of DNA methylation data, which were included in ARIES, was obtained from the MRC, the Economic and Social Research Council, the National Institutes of Health, and other sources. ARIES is maintained under the auspices of the MRC Integrative Epidemiology Unit at the University of Bristol (grants MC_UU_12013/2 and MC_UU_12013/8). A comprehensive list of funding grants is available on the ALSPAC website (http://www.bristol.ac.uk/alspac/external/documents/grant-acknowledgements.pdf).

The content of this article is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The funders played no role in the design, execution, analysis, or interpretation of the data or in the writing of the manuscript. This publication is the work of the authors, each of whom serves as a guarantor for the contents of this paper.

We are extremely grateful to all of the families who took part in the ALSPAC Study, the midwives for their help in recruiting them, and the whole ALSPAC team, which

includes interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists, and nurses.

This paper was partly presented in poster form at the 35th Annual Meeting of the International Society for Traumatic Stress Studies, Boston, Massachusetts, November 14–16, 2019. It was also presented as a poster at the 53rd Annual Meeting of the Society for Epidemiologic Research, held virtually on December 16–18, 2020.

Conflict of interest: none declared.

REFERENCES

1. De Stavola BL, Nitsch D, dos Santos Silva I, et al. Statistical issues in life course epidemiology. *Am J Epidemiol*. 2006; 163(1):84–96.
2. Ben-Shlomo Y, Cooper R, Kuh D. The last two decades of life course epidemiology, and its relevance for research on ageing. *Int J Epidemiol*. 2016;45(4):973–988.
3. Kuh D, Cooper R, Hardy R, et al., eds. *A Life Course Approach to Healthy Ageing*. Oxford, United Kingdom: Oxford University Press; 2013.
4. 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(2):285–293.
5. Kuh D. *A Life Course Approach to Chronic Disease Epidemiology*. Oxford, United Kingdom: Oxford University Press; 1997:344.
6. Koenen KC, Rudenstine S, Susser E, et al., eds. *A Life Course Approach to Mental Disorders*. Oxford, United Kingdom: Oxford University Press; 2013.
7. Mishra G, Nitsch D, Black S, et al. A structured approach to modelling the effects of binary exposure variables over the life course. *Int J Epidemiol*. 2009;38(2):528–537.
8. Efron B, Hastie T, Johnstone I, et al. Least angle regression. *Ann Stat*. 2004;32(2):407–499.
9. Smith ADAC, Heron J, Mishra G, et al. Model selection of the effect of binary exposures over the life course. *Epidemiology*. 2015;26(5):719–726.
10. Smith ADAC, Hardy R, Heron J, et al. A structured approach to hypotheses involving continuous exposures over the life course. *Int J Epidemiol*. 2016;45(4):1271–1279.
11. Madathil S, Joseph L, Hardy R, et al. A Bayesian approach to investigate life course hypotheses involving continuous exposures. *Int J Epidemiol*. 2018;47(5):1623–1635.
12. Dunn EC, Soare TW, Raffeld MR, et al. What life course theoretical models best explain the relationship between exposure to childhood adversity and psychopathology symptoms: recency, accumulation, or sensitive periods? *Psychol Med*. 2018;48(15):2562–2572.
13. Cooper R, Mishra GD, Kuh D. Physical activity across adulthood and physical performance in midlife: findings from a British birth cohort. *Am J Prev Med*. 2011;41(4):376–384.
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(6): 632–640.
15. 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(5):655–660.
16. Bann D, Kuh D, Wills AK, et al. Physical activity across adulthood in relation to fat and lean body mass in early old age: findings from the Medical Research Council National Survey of Health and Development, 1946–2010. *Am J Epidemiol*. 2014;179(10):1197–1207.
17. Dunn EC, Crawford KM, Soare TW, et al. Exposure to childhood adversity and deficits in emotion recognition: results from a large, population-based sample. *J Child Psychol Psychiatry*. 2018;59(8):845–854.
18. Nicolau B, Madathil SA, Castonguay G, et al. Shared social mechanisms underlying the risk of nine cancers: a life course study. *Int J Cancer*. 2019;144(1):59–67.
19. Huang JY, Gavin AR, Richardson TS, et al. Accounting for life-course exposures in epigenetic biomarker association studies: early life socioeconomic position, candidate gene DNA methylation, and adult cardiometabolic risk. *Am J Epidemiol*. 2016;184(7):520–531.
20. Hughes A, Smart M, Gorrie-Stone T, et al. Socioeconomic position and DNA methylation age acceleration across the life course. *Am J Epidemiol*. 2018;187(11):2346–2354.
21. Everson TM, Marsit CJ. Integrating -omics approaches into human population-based studies of prenatal and early-life exposures. *Curr Environ Health Rep*. 2018;5(3): 328–337.
22. Visscher PM, Wray NR, Zhang Q, et al. 10 years of GWAS discovery: biology, function, and translation. *Am J Hum Genet*. 2017;101(1):5–22.
23. Dunn EC, Soare TW, Zhu Y, et al. Sensitive periods for the effect of childhood adversity on DNA methylation: results from a prospective, longitudinal study. *Biol Psychiatry*. 2019; 85(10):838–849.
24. Amrhein V, Greenland S, McShane B. Scientists rise up against statistical significance. *Nature*. 2019;567(7748): 305–307.
25. Wasserstein RL, Schirm AL, Lazar NA. Moving to a world beyond "$p < 0.05$". *Am Stat*. 2019;73(suppl 1):1–19.
26. Buja A, Brown L. Discussion: a significance test for the lasso. *Ann Stat*. 2014;42(2):509–517.
27. Lockhart R, Taylor J, Tibshirani RJ, et al. A significance test for the lasso. *Ann Stat*. 2014;42(2):413–468.
28. Lockhart R, Taylor J, Tibshirani R, et al. Package 'covTest'. (R package, version 1.02). https://cran.r-project.org/web/packages/covTest/index.html. Published August 14, 2013. Accessed January 23, 2019.
29. Tibshirani RJ, Taylor J, Lockhart R, et al. Exact post-selection inference for sequential regression procedures. *J Am Stat Assoc*. 2016;111(514):600–620.
30. Tibshirani R, Tibshirani R, Taylor J, et al. Package 'selectiveInference'. (R package, version 1.2.0). https://cran.r-project.org/web/packages/selectiveInference/selectiveInference.pdf. Published September 7, 2019. Accessed October 4, 2017.
31. thedunnlab. simulations. https://github.com/thedunnlab/simulations. Published October 11, 2019. Accessed March 16, 2021.
32. Tsai P-C, Bell JT. Power and sample size estimation for epigenome-wide association scans to detect differential DNA methylation. *Int J Epidemiol*. 2015;44(4): 1429–1441.
33. 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(8):2201–2217.
34. Sharp GC, Lawlor DA, Richmond RC, et al. Maternal pre-pregnancy BMI and gestational weight gain, offspring

DNA methylation and later offspring adiposity: findings from the Avon Longitudinal Study of Parents and Children. *Int J Epidemiol*. 2015;44(4):1288–1304.
35. Frisch R, Waugh FV. Partial time regressions as compared with individual trends. *Econometrica*. 1933;1(4):387–401.
36. Lovell MC. Seasonal adjustment of economic time series and multiple regression analysis. *J Am Stat Assoc*. 1963;58(304): 993–1010.
37. Yamada H. The Frisch–Waugh–Lovell theorem for the lasso and the ridge regression. *Commun Stat Theory Methods*. 2017;46(21):10897–10902.
38. Khoury MJ. A primer series on -omic technologies for the practice of epidemiology. *Am J Epidemiol*. 2014;180(2): 127–128.
39. Khoury MJ. Planning for the future of epidemiology in the era of big data and precision medicine. *Am J Epidemiol*. 2015;182(12):977–979.
40. Kuller LH. Epidemiologists of the future: data collectors or scientists? *Am J Epidemiol*. 2019;188(5):890–895.
41. König IR. Presidential address: six open questions to genetic epidemiologists. *Genet Epidemiol*. 2019;43(3): 242–249.
42. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. *Stat Med*. 2019;38(11): 2074–2102.
43. Lawlor DA, Tilling K, Davey SG. Triangulation in aetiological epidemiology. *Int J Epidemiol*. 2016;45(6): 1866–1886.
44. Howe LD, Smith AD, MacDonald-Wallis C, et al. Relationship between mediation analysis and the structured life course approach. *Int J Epidemiol*. 2016;45(4): 1280–1294.
45. Robins JM, Hernán MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology*. 2000;11(5):550–560.
46. Cole SR, Hernán MA. Constructing inverse probability weights for marginal structural models. *Am J Epidemiol*. 2008;168(6):656–664.
47. Kim S, Schliekelman P. Prioritizing hypothesis tests for high throughput data. *Bioinformatics*. 2016;32(6):850–858.
48. Iotchkova V, Ritchie GRS, Geihs M, et al. GARFIELD classifies disease-relevant genomic features through integration of functional annotations with association signals. *Nat Genet*. 2019;51(2):343–353.

A Structured Approach to Evaluating Life-Course Hypotheses: Moving Beyond Analyses of Exposed Versus Unexposed in the -Omics Context

Yiwen Zhu, Andrew J. Simpkin, Matthew J. Suderman, Alexandre A. Lussier, Esther Walton, Erin C. Dunn§, and Andrew D.A.C. Smith§
§Both senior authors contributed equally to this work. Their names appear alphabetically.

Contents

1 Web Appendix 1

This section provides details on the statistical methods examined in the current study. We introduce the regression setup formally, followed by an overview of two variable selection procedures and five methods for making statistical inference in the structured life course modeling approach (SLCMA), which were assessed in the current study. Details on a new confidence interval calculation for the max-$|t|$-test are also provided. A summary of the technical details is provided in **Web Table 3** for quick reference.

For a sample of size n, let $\boldsymbol{y}$ be the vector of responses and $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p$ be vectors of p predictors. These are assumed to be centered and standardized such that $\sum_{i=1}^{n} y_i = 0$, and $\sum_{i=1}^{n} x_{ij} = 0$ and $\sum_{i=1}^{n} x_{ij}^2 = 1$ for all $j = 1, \ldots, p$. The response $\boldsymbol{y}$ is assumed to be a realization of the random vector generated by the model

$$\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1}$$

where $X = (\boldsymbol{x}_1 \cdots \boldsymbol{x}_p)$ and $\boldsymbol{\varepsilon} \sim \mathrm{N}(0, \sigma^2 I)$. The predictor that explains the most variation in the response is the predictor with the largest correlation with the response, i.e. the predictor that maximizes $|\boldsymbol{x}_j^T \boldsymbol{y}|$.

1.1 Variable selection procedures

Two variable selection procedures that find the single predictor that explains the most variation in the outcome (in their first step) are forward stepwise regression and least angle regression (LARS). We therefore considered post-selection inference methods that were developed for these two procedures. This section contains a short overview of these two procedures.

Forward stepwise regression fits a sequence of models with an increasing number of predictors. At each step, the procedure selects the predictor not already in the model that has the largest correlation with the residuals obtained from the current model. At the first step, there are no predictors in the model, the residuals are therefore simply $\boldsymbol{y}$, and the correlations with the residuals are contained in $\boldsymbol{r}$ Hence the first-selected predictor maximizes $|\boldsymbol{x}_j^T \boldsymbol{y}|$.

LARS [7] is related to the lasso [19]. The lasso estimate $\hat{\boldsymbol{\beta}}$ minimizes $\frac{1}{2}||\boldsymbol{y} - X\boldsymbol{\beta}||_2^2 + \lambda||\boldsymbol{\beta}||_1$ for some fixed positive value of the smoothing parameter λ. For sufficiently large λ we have $\hat{\boldsymbol{\beta}} = \boldsymbol{0}$, i.e. the lasso has selected a model with no predictors. As λ decreases the model selected by the lasso increases in complexity. LARS is a procedure for identifying the sequence of lasso models and estimates produced as λ decreases. The first step of LARS identifies the value λ_1 below which the lasso selects its first predictor, the second step identifies the value λ_2 below which a second predictor is selected. The predictor selected at λ_1 is that which maximizes $|\boldsymbol{x}_j^T \boldsymbol{y}|$, as in forward stepwise regression. This second predictor is not necessarily the same as that selected in the second step of forward stepwise regression.

For simplicity of notation, we will assume that $\boldsymbol{x}_1$ is the predictor selected in the first step of both forward stepwise regression and LARS. The model containing only the first-selected predictor simplifies to

$$\boldsymbol{Y} = \boldsymbol{x}_1 \beta_1 + \boldsymbol{\varepsilon}. \tag{2}$$

Note that r_1 is the ordinary least squares estimate for the regression coefficient β_1 in this model.

1.2 Post-selection inference methods

This section gives an overview of methods for calculating P values and confidence intervals for the regression coefficient of the first-selected predictor.

1.2.1 Naïve calculations

A typical implementation of forward stepwise regression ignores the selective nature of the procedure. Inference is essentially based on the test statistic for the first selected predictor, ignoring the fact that this predictor was selected due to its having the largest correlation with the residuals, which would artificially give it a test statistic larger than that expected under the null hypothesis.

In the context of the first predictor selected by forward stepwise regression, this naïve method of inference would involve fitting the simple linear regression model in (2) and testing the hypothesis $H_0 : \beta_1 = 0$ against a two-sided alternative. If H_0 is rejected at the α level of significance, then the probability of making a type I error will be potentially much larger than α. This is because the usual hypothesis test assumes that $\boldsymbol{x}_1$ has been selected from the predictors at random, rather than selected because it has the largest correlation with the response (and hence largest standardized regression coefficient). For $p = 10$ and $\alpha = 5\%$, the probability of a type I error would be approximately 39% using this method [12].

1.2.2 Bonferroni correction

It is possible to control the type I error in the naïve approach by means of a Bonferroni correction, dividing the significance level α by the number of predictors, p (or equivalently, multiplying the P value by the number of predictors and capping at 1). The resulting probability of type I error would be less than α. However, this would be a very conservative approach, as the Bonferroni correction assumes that the regression coefficients are uncorrelated. In practice, this will only occur if the predictors are orthogonal. In the case of general predictors, Bonferroni correction will result in a loss of statistical power. The methods in the following sections attempt to control the probability of type I error without loss of statistical power, by further making use of the correlation between predictors.

1.2.3 Covariance test

Lockhart et al. [12] developed the covariance test as "a significance test for the lasso". It provides a P value for the selected variable that takes into account the selective nature of the sequence of lasso models.

For the first predictor selected by LARS, the null hypothesis considered by the covariance test is $H_0 : \boldsymbol{\beta} = \mathbf{0}$, and the test statistic is

$$\lambda_1(\lambda_1 - \lambda_2)/\sigma^2. \tag{3}$$

If σ^2 is known this test statistic will have, asymptotically, an Exp(1) distribution under the null hypothesis. Since λ_1 does not depend on the correlation between a particular predictor and the response, but on the maximum correlation between predictor and response, this test statistic takes into account the fact that LARS has not selected a predictor at random, but selected the predictor with the largest correlation with the response.

Lockhart et al. [12] demonstrated the distribution of T_1 with an example with $n = 100$ and $p = 10$ orthogonal predictors. The authors claim that, in their example and under the null hypothesis, the quantiles of T_1 were "decently matched" to those of an Exp(1) distribution.

If the variance σ^2 is unknown it can be replaced by an estimate. Provided that $n > p$, the variance can be estimated by fitting the full linear model (1). When an estimated variance is used, the covariance test statistic will have, asymptotically, an $\mathrm{F}(2, n-p)$ distribution under the null hypothesis. Further details regarding variance estimation for post-selection inference is discussed in Reid et al. [15].

The covariance test does not directly yield a confidence interval for the regression coefficient β_1. Smith et al. [18] proposed a method for modifying the usual confidence interval to account for the selective nature of the model being presented, based on the covariance test P value. They showed using simulations that 95% confidence intervals calculated this way had 95% coverage in a typical structured approach application. However, the usual confidence interval and the covariance test P value are not based on a common set of statistics, and as a result the confidence intervals of Smith et al. [18] and the covariance test P values are not consistent. That is, the 95% confidence interval may contain 0 even if the P value is less than 5%, and vice versa. This can cause confusion if, as is typical in many applications, confidence intervals and P values are displayed side by side in results.

1.2.4 Selective inference

Tibshirani et al. [21] proposed a new set of inference tools applicable to forward stepwise regression and LARS, which are available in the `selectiveInference` R package [20]. The authors identified variable selection procedures that made estimates under polyhedral constraints. As a result, P values and confidence intervals are calculated based on a truncated Gaussian distribution.

For the first step of the variable selection procedure, the standard implementation of the `selectiveInference` package calculates a P value pertaining to the null hypothesis $H_0 : \beta_1 = 0$. The P value is the probability that the estimated regression coefficient would be more extreme than r_1, under H_0 and conditional on the fact that $\boldsymbol{x}_1$ is the first-selected predictor and that the estimated regression coefficient has the same sign as r_1. This P value can be shown to be

$$\frac{1 - \Phi(\lambda_1/\sigma)}{1 - \Phi(\lambda_2/\sigma)} \tag{4}$$

where Φ is the cumulative distribution function for a standard normal distribution.

In its standard implementation the P values and confidence intervals calculated by the `selectiveInference` package are not consistent. The 95% confidence intervals will contain 0 if and only if the corresponding P value is greater than 2.5%, not 5%. The reason for this is the P value in (4) is effectively that of a one-sided test, due to conditioning on the regression coefficient having the same sign as r_1. Thus a confidence interval would have to be one-sided to be consistent with the P value in (4).

1.2.5 max-$|t|$-test

Buja and Brown [4] proposed a hypothesis test for forward selection, based on the largest t-value of all predictors not yet included in the model at a certain step. We present basic

details of this hypothesis test at the first step of the procedure, and a novel method for calculating consistent confidence intervals for the regression coefficient in the first-selected model.

As the predictors $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p$ share a common scale, the t-values in the first step of the procedure will be proportional to the correlations between the predictors and the response. Therefore we can use the largest correlation as a test statistic. Let $\boldsymbol{r} = X^T\boldsymbol{y}$ be the vector of observed correlations, and let $\boldsymbol{R} = X^T\boldsymbol{Y}$, so that $\boldsymbol{R} \sim \mathrm{N}(X^T\boldsymbol{\beta}, \sigma^2 X^T X)$ under model (1). As we have assumed that $\boldsymbol{x}_1$ is the first predictor selected by LARS and forward selection, as it has the largest correlation with $\boldsymbol{y}$, then r_1 is the observed value of the test statistic, and $|r_1| = \max_j |r_j|$.

We consider the hypothesis $H_0 : \boldsymbol{\beta} = \mathbf{0}$ versus $H_A : \boldsymbol{\beta} \neq \mathbf{0}$. The P value for the max-$|t|$-test equals

$$\begin{aligned} & P_0\Big(|R_1| > |r_1| \,\Big|\, |R_1| = \max_j |R_j|\Big) \\ = \;& 1 - P_0\big(-|r_1| \leq R_1 \leq |r_1| \cap \cdots \cap -|r_1| \leq R_p \leq |r_1|\big). \end{aligned} \tag{5}$$

where P_0 refers to the probability under H_0.

Hence the P value is the probability that $\boldsymbol{R}$ lies outside a cube of radius $|r_1|$. Under H_0, the distribution of $\boldsymbol{R}$ is $\mathrm{N}(\mathbf{0}, \sigma^2 X^T X)$. The probability in (5) can be calculated using existing software for the multivariate normal distribution if σ^2 is known. If σ^2 is unknown, we can estimate it from the residuals of the full linear model as for the covariance test or selective Inference package and calculate the P value from a multivariate t-distribution with $n - p$ degrees of freedom.

Having selected the model in (2) we can construct a 95% confidence interval for β_1 that is consistent with the P value in (5). This confidence interval is the set of all β values that would give a P value not less than 0.05 when testing $H_0 : \beta_1 = \beta$ against a two-sided alternative. Under the model in (2), we have $E(R_1) = \beta_1$, so a suitable test statistic would be $R_1 - \beta$. A 95% confidence interval is given by

$$\begin{aligned} & \Big\{\beta : P_\beta\Big(|R_1 - \beta| > |r_1 - \beta| \,\Big|\, |R_1| = \max_j |R_j|\Big) \geq 0.05\Big\} \\ = \;& \Big\{\beta : P_\beta\Big(|R_1 - \beta| \leq |r_1 - \beta| \,\Big|\, |R_1| = \max_j |R_j|\Big)) \leq 0.95\Big\}, \end{aligned}$$

where P_β refers to the probability under $H_0 : \beta_1 = \beta$. The limits of this interval must be found using numerical methods. To calculate whether a certain value of β belongs inside the interval requires calculation of the form

$$\begin{aligned} & P_\beta\Big(R_1 \leq r \,\Big|\, |R_1| = \max_j R_j\Big) \\ = \;& \frac{P_\beta\big(R_1 \leq r \cap |R_1| = \max_j |R_j|\big)}{P_\beta\big(R_1 \leq \infty \cap |R_1| = \max_j |R_j|\big)} \end{aligned} \tag{6}$$

for $r = \beta \pm |r_1 - \beta|$.

Under H_0 we have $\boldsymbol{R} \sim \mathrm{N}(X^T\boldsymbol{x}_1\beta, \sigma^2 X^T X)$. If σ^2 is estimated then a multivariate t-distribution with $n - p$ degrees of freedom should be used for calculation instead of a multivariate normal distribution. Existing software for multivariate normal and t distributions can calculate probabilities over (potentially infinite) cuboid regions. However, probabilities

of the form encountered in (6) require calculation over non-cuboid regions. A simple linear transformation allows these probabilities to be calculated using existing software.

The probabilities in (6) can be written as follows

$$\begin{aligned} & P_\beta\big(R_1 \le r \cap |R_1| = \max_j |R_j|\big) \\ = & \begin{cases} 1 - P_\beta\big(R_1 \ge r \cap |R_1| = \max_j |R_j|\big) & r \ge 0 \\ P_{-\beta}\big(R_1 \ge -r \cap |R_1| = \max_j |R_j|\big) & r < 0. \end{cases} \end{aligned}$$

We will discuss how to calculate a general probability

$$P\Big(R_1 \ge r \cap |R_1| = \max_j |R_j| \,\Big|\, \boldsymbol{R} \sim \mathrm{N}(\boldsymbol{\mu}, \Sigma)\Big) \tag{7}$$

for $r \ge 0$. Note that

$$\begin{aligned} & R_1 \ge r \cap |R_1| = \max_j |R_j| \\ = & R_1 \ge r \cap -R_1 \le R_2 \le R_1 \cap \cdots \cap -R_1 \le R_p \le R_1 \end{aligned}$$

and this set of inequalities is equivalent to the intersection of the following set of inequalities:

$$\begin{aligned} R_1 &\ge r \\ R_1 - R_2 &\ge 0 \\ R_1 + R_2 &\ge 0 \\ &\vdots \\ R_1 - R_p &\ge 0 \\ R_1 + R_p &\ge 0. \end{aligned} \tag{8}$$

Let C be a $2p - 1 \times p$ matrix with

$$C_{i,j} = \begin{cases} 1 & j = 1 \\ -1 & i = j \text{ even} \\ 1 & i = j \text{ odd} \\ 0 & \text{otherwise} \end{cases}$$

Then the inequalities in (8) are satisfied by $C\boldsymbol{R} \ge \boldsymbol{r}'$ where $\boldsymbol{r}' = (r, 0, \ldots, 0)^T$. As C is a linear transformation and $\boldsymbol{R}$ has either a multivariate normal or multivariate t distribution, then $\boldsymbol{R}' = C\boldsymbol{R}$ will have a multivariate normal or t distribution. Hence the probability in (7) is equal to

$$P\Big(\boldsymbol{R}' \ge \boldsymbol{r}' \,\Big|\, \boldsymbol{R}' \sim \mathrm{N}(C\boldsymbol{\mu}, C\Sigma C^T)\Big)$$

and this can be calculated using existing software.

1.3 Simulations setup and data generating process

We included a brief description of the simulations setup in the main text. Here we present full details on the data generating process. We leveraged observed data from the empirical example such that the simulation analyses closely resembled a real-world example of SLCMA application in omics.

1.3.1 Scenario 1: normal outcomes

In the first scenario, the seven exposure variables were resampled from observed data, consisting of five sensitive periods (binary variables taking the value of 0 or 1), accumulation (sum of all five sensitive periods, ranging from 0 to 5), and recency (a weighted sum, with weights defined as age at assessment). The j^{th} outcome (i.e., DNA methylation, $j = 1, \cdots, 485000$) was simulated from a standard normal distribution, $y_j \sim \mathcal{N}(0, 1)$. Because DNA methylation values ('beta' values) may not following a normal distribution, we considered the consequence of having a non-normally distributed outcome in the second scenario. Considering a normally distributed outcome was still useful, as it would help illustrate the performance of the methods when the assumption held.

To assess FWER under the null hypothesis, we ran a single simulation of 485,000 tests and examined the distributions of observed P values against their expected distribution. To assess statistical power and CI coverage, we ran simulations in which the outcome variable was correlated with one of the predictors and then varied the correlation between outcome and predictor such that the variance explained by the predictor r^2 varied from 0.01 to 0.1. When only the first sensitive period hypothesis, denoted by X_1, was simulated to be associated with the outcome, we generated the j^{th} outcome as follows:

$$y_j = X_{1,i}\beta_j + \epsilon_{ij}, \text{where } \beta_j = \sqrt{\frac{r^2}{1-r^2}}, \epsilon_{ij} \sim \mathcal{N}(0, 1)$$

1.3.2 Scenario 2: empirical outcomes

We aimed to consider non-normally distributed outcomes that closely resemble observed DNA methylation values. The simulation of the exposures was identical to the process described in scenario 1. To assess the FWER under the null hypothesis, we resampled the real predictors and DNAm values from ALSPAC separately. The resampling breaks the predictor-outcome link and hence removes any observed association between the two, while maintaining the empirical distributions of DNAm. In the assessment of statistical power and confidence interval coverage, outcomes were simulated to follow beta distributions and effect sizes were parameterized as the difference in mean levels of DNAm between the exposed and unexposed at the first sensitive period (ΔDNAm), ranging from 0.05 to 0.5.

The number of tests and P value threshold were the same as Scenario 1. We additionally considered a transformation of the DNAm values from beta values (y) to M values equivalent to $M = \log_2 \frac{y}{1-y}$, which are sometimes used to stabilize variance [5].

1.4 Discussion

In addition to the discussion provided in the main text, we would like to highlight a few technical details that may also influence one's preference for one post-selection inference method over another. First of all, the consistency of the confidence interval (CI) and the corresponding P value may make the max-$|t|$-test more favorable. While the selective inference method provided desired confidence interval coverage, the confidence intervals and P values calculated are not consistent: the confidence intervals are two-sided but the P value effectively tests a one-sided hypothesis. Tibshirani et al. [21] argued in favor of this inconsistency, giving the reasons that the one-sided P value would be expected to have more statistical

power, while practitioners would prefer to report two-sided confidence intervals. We would further argue that practitioners would prefer to report confidence intervals consistent with observed P values, given that both are frequently reported side by side. Consistent confidence intervals can be provided by the max-$|t|$-test introduced in this paper.

Second, when considering a compound hypothesis in the simulations, we noticed that the statistical power of the selective inference was reduced. As has been noted by Fan and Ke [8] and Bühlmann et al. [3], if there is a second predictor with a non-zero regression coefficient, then λ_2 will be closer to λ_1 and the covariance test statistic will be smaller than if there were no such second predictor. We further note that the selective inference P value (4) will be larger under this scenario. Hence the statistical power of the covariance test and selective inference method may be severely reduced if another predictor also has a large contribution to the outcome variation, even when only considering inference regarding the first-selected predictor. There is no theoretical basis for such a reduction in power when using the max-$|t|$-test, which was consistent with our observation in the simulations. We recommend that practitioners conduct their own simulations to determine statistical power if there is any uncertainty on this point.

Third, post-selection inference methods are available for generalized linear models. Although implementations of the covariance test were available in the `covTest` R package [11], these are no longer recommended by the package authors. The `selectiveInference` package can be used for binary or Cox regression [20], but further simulation is required to confirm its suitability in high-throughput applications. The fact that further Bonferroni correction did not result in a significant loss of statistical power indicates that this conservative method could potentially be used if post-selection software is not available for certain nonlinear regression models.

1.5 Estimating family-wise error rate (FWER)

To further investigate the FWER, we performed repeated simulation experiments under a theoretical scenario. Specifically, we based the setup on the simulations described by Lockhart et al. [12], who used a simulated example to investigate the distribution of the covariance test statistic. We ran 2 000 simulation experiments for each set of parameters to allow the confidence interval of the FWER to have a radius of 1%, setting α to 5%. In each of the 2 000 simulations, we simulated a sample size of $n = 100$ and $p = 10$ uncorrelated predictors with a Gaussian distribution. The response was also generated from a Gaussian distribution. We set $m = 1$, as in Lockhart et al. [12], but also investigated values of m =10, 100, or 1 000 to assess how calculations were affected by the number of tests. The residual variance σ^2 was considered fixed and known, hence the covariance test statistic was considered against an Exp(1) distribution, and a multivariate normal distribution was used for the max-$|t|$-test.

The estimates of FWER for varying numbers of tests performed are presented in **Web Table 4**. As predicted by Lockhart et al. [12], the FWER for the naïve method was not significantly different from 39%, no matter how many tests were performed. In contrast, the conservative Bonferroni correction (using an individual test significance level of $5/pm\%$) gave a FWER that was not significantly different from 5% for all considered values of m. In this scenario, Bonferroni correction is not overly conservative as the predictors are uncorrelated. Hence the p tests of regression coefficients that are implicitly considered during variable selection are independent.

The selective inference method and the max-$|t|$-test gave FWER that were not significantly different from 5% for all considered values of m. The covariance test gave a FWER of approximately 5% for $m = 1$, but for increasing m the FWER increased. For m=1 000 the covariance test FWER was similar to that of the naive method. We took this to indicate that, below 0.05, the P values generated by the covariance test under the null hypothesis are smaller than expected.

The conclusions drawn from this set of simulations are consistent with the conclusions drawn from **Figure 1**.

2 Web Appendix 2

The follow R code shows how the P values and the confidence intervals of the post-selection inference methods compared in this study can be computed in R, where `X_hypos` is the design matrix, `y` the outcome, and `npred` the number of predictors. The code is also available on GitHub: `https://github.com/thedunnlab/simulations`

```
library(lars)
# archived version of the covTest package can be retrieved here:
# https://cran.r-project.org/src/contrib/Archive/covTest/
library(covTest)
library(selectiveInference)
library(mvtnorm)


## X_hypos: a matrix of the predictors
## y: outcome
## npred: number of predictors
## n: sample size


#### functions for confidence interval for the max-|t|-test ---

# Calculates the probability in (5)
Psi <- function(z, p, mu, df, s2, Corr) {
  C <- rbind(diag(p),-diag(p))
  C <- C[-(p+1),]
  C[,1] <- 1
  pmvt(lower=c(z, rep(0,2*p-2)), upper=rep(Inf,2*p-1),
       delta=as.vector(C %*% mu), df=df, sigma=s2*C %*% Corr %*% t(C), type="shifted")
}


# Calculates the probability in (4)
Pconditional <- function(r, largest, mu, df, s2, Corr) {
  # Reorder so that variable in position 1 is the first one selected
  p <- length(mu)
  mu <- mu[c(largest, (1:p)[-largest])]
  Corr <- Corr[c(largest, (1:p)[-largest]),]
  Corr <- Corr[,c(largest, (1:p)[-largest])]
  # Calculate denominator in (4)
  lower.denom <- Psi(0, p, -mu, df, s2, Corr)
  upper.denom <- Psi(0, p,  mu, df, s2, Corr)
  # Calculate numerator in (4), according to page x
  if(r >= 0) {
    numer <- Psi( r, p,  mu, df, s2, Corr)
    prob <- 1 - numer / (lower.denom + upper.denom)
  } else {
    numer <- Psi(-r, p, -mu, df, s2, Corr)
    prob <- numer / (lower.denom + upper.denom)
```

```
  }
  prob
}

Paccept <- function(beta) {
  Pconditional(beta+abs(Xty[selection]-beta), selection,
               XtX[,selection] * beta, n-7,  s^2, XtX) -
    Pconditional(beta-abs(Xty[selection]-beta), selection,
                 XtX[,selection] * beta, n-7,  s^2, XtX) - 0.95
}

#### Run SLCMA ----

# Normalize the design matrix
col_mean <- apply(X_hypos, 2, mean)
X_centered <- X_hypos - rep(col_mean, rep(n, npred)) #subtract mean
col_sss <- apply(X_centered, 2, function(x) sqrt(sum(x^2)))
X_normed <- X_centered / rep(col_sss, rep(n, npred)) #divide by sqrt sum squares

Xt <- t(X_normed)
XtX <- Xt %*% X_normed
Xty <- Xt %*% y

y_centered <- y - mean(y)

## select the predictor with the highest correlation
selection <- which.max(abs(Xty))

## fit OLS
coeftable <- summary(lm(y ~ X_normed[,selection]))$coef


## Naive calculations ----
p.naive <- coeftable[2,4]
lower.naive <- coeftable[2,1] + qt(0.025, n-n_hypo)*coeftable[2,2]
upper.naive <- coeftable[2,1] + qt(0.975, n-n_hypo)*coeftable[2,2]

## Naive calculations + Bonferroni correction ----
p.bonf <- ifelse(p.naive*npred <= 1, p.naive*npred, 1)


## Covariance test ----
lasso <- lars(X_hypos, y)
tt <- covTest(lasso,X_hypos,sigma.est=1,y,maxp=2)$results[1,2]
p.covTest <- 1 - pexp(tt, 1)
# Code from Smith et al. (2015)
thep <- p.covTest/2
lower.covTest <- -1
upper.covTest <- 1
if(thep < 0.05) {
  lower.covTest <- coeftable[2,1]+qnorm((0.025-thep/2)/(1-thep))*coeftable[2,2]
```

```
  upper.covTest <- coeftable[2,1]+qnorm((0.975-thep/2)/(1-thep))*coeftable[2,2]
}
if(lower.covTest <= 0 & upper.covTest >= 0 & thep < 0.975) {
  lower.covTest <- coeftable[2,1]+qnorm(0.025/(1-thep))*coeftable[2,2]
  upper.covTest <- coeftable[2,1]+qnorm((0.975-thep)/(1-thep))*coeftable[2,2]
}
if(thep >= 0.975) {
  lower.covTest <- 0
  upper.covTest <- 0
}

## Selective inference ----
larfit <- lar(X_normed, y, maxsteps=3)
inference <- larInf(larfit, type="active", alpha=0.05)
p.sI <- inference$pv[1]
lower.sI <- inference$ci[1,1]
upper.sI <- inference$ci[1,2]

## Max-|t| test ----
absbeta <- abs(Xty[selection])
s <- summary(lm(y_centered ~ X_normed))$sigma
p.maxt <- 1 -
  pmvt(lower=-rep(absbeta,npred),
       upper= rep(absbeta,npred),
       delta= rep(0,npred),
       df= n-npred, sigma= s^2 * XtX)

search_middle <- Xty[selection]
search_radius <- 3*s*XtX[selection,selection]
# lower limit
lower.maxt <- uniroot(Paccept,
        lower=search_middle-search_radius, upper=search_middle)$root
# upper limit
upper.maxt <- uniroot(Paccept,
        lower=search_middle, upper=search_middle+search_radius)$root
```

3 Web Appendix 3

3.1 Sample and procedure

The empirical exposure and outcome data used in our simulations came from the Avon Longitudinal Study of Parents and Children (ALSPAC), a prospective, longitudinal birth cohort of children born to mothers living in the county of Avon, England (120 miles west of London) with estimated delivery dates between April 1991 and December 1992 [9, 2]. Approximately 85 percent of eligible pregnant women agreed to participate (N=14,541), and 99% of eligible live births (n=14,062) who were alive at one year of age (n=13,988 children) were enrolled. Response rates to data collection have been good (75% have completed at least one follow-up). Informed consent for the use of data collected via questionnaires and clinics was obtained from participants following the recommendations of the ALSPAC Ethics and Law Committee at the time. Ethical approval for the study was obtained from the ALSPAC Ethics and Law Committee and the Local Research Ethics Committees. Consent for biological samples has been collected in accordance with the Human Tissue Act 2004 [1]. More details are available on the ALSPAC website, including a fully searchable data dictionary: `http://www.bristol.ac.uk/alspac/researchers/our-data/`. The ALSPAC generated blood-based DNAm profiles at 7 years of age as part of the Accessible Resource for Integrated Epigenomic Studies (ARIES), a subsample of 1018 mother–child pairs from the ALSPAC. The ARIES mother-child pairs were randomly selected out of those with complete data across at least five waves of data collection [16].

3.2 Measures

We used data capturing the exposure to sexual or physical abuse (by anyone) and constructed the following hypotheses: five sensitive periods (at ages 1.5 years, 2.5 years, 3.5 years, 4.75 years, 5.75 years, and 6.75 years), accumulation, and recency. The first sensitive period hypothesis was set to be the true underlying hypothesis in the power and confidence interval coverage simulations. The prevalence of the exposure at the six time points ranged from 2.61% to 3.96%. Exposure to sexual or physical abuse was determined through an item asking the mother to indicate whether or not the child had been exposed to either sexual or physical abuse from anyone at each of the six time points listed above. Reports of sexual or physical abuse were not reported to child welfare agencies. Other available types of exposure to childhood adversity in ALSPAC are described by Dunn et al. [6].

DNAm was measured at 485,000 CpG dinucleotide sites across the genome using the Illumina Infinium Human Methylation 450K BeadChip microarray. DNA for this assay was extracted from peripheral blood leukocytes at age 7. DNAm levels are expressed as a 'beta' value, representing the proportion of cells methylated at each interrogated CpG site. Detailed descriptions of the preprocessing and quality control procedures are provided elsewhere [6, 16].

The covariates included in the empirical analyses were consistent with the adjustment by Dunn et al. [6]. They were: child sex, child race and/or ethnicity; child birth weight; maternal age; number of previous pregnancies; sustained maternal smoking during pregnancy; and parent social class.

3.3 Adjusting for covariates

While we focused on assessing relationships between exposures capturing life course theories and omics outcomes in the simulations, the associations are usually confounded by other factors in practice, such as socioeconomic status or maternal smoking status during pregnancy [6]. In previous studies, adjustment had been formerly done by regressing exposures on the covariates and using the residuals of the exposures in the SLCMA [18].

An alternative method that can be used to adjust for covariates in a linear regression setting is to apply the Frisch-Waugh-Lovell (FWL) theorem, or partitioned regression [10, 13]. The method has been proven to yield the same regression coefficients and residual variance as a fully adjusted model [14]. The FWL theorem was first proposed by two econometricians, Frisch and Waugh [10], to highlight a useful property of ordinary least squares such that a two-step approach to detrend the independent and dependent variables yields the same regression coefficients as a fully adjusted regression model with the trend variables included as covariates. Lovell showed that the adjustment remains true for any nonempty subset of explanatory variables (i.e., it does not just apply to trend variables) [13]. The proof of this theorem can be found in several previous publications [10, 13, 14]. It has since been proven in the context of penalized regression as well, such as the lasso or ridge regression [22].

However, it remained unclear whether this theorem would be applicable to post-selection inference methods (such as selective inference or the max-$|t|$-test) and whether additionally regressing the outcomes on the covariates would result in smaller residual variances and larger test statistics. Therefore, we assessed the FWER in a similar manner as in scenario 2 presented in the main results, by running one simulation experiment with resampled empirical outcomes (n=700). As seen in **Web Figure 5**, the P value distributions were similar to what we observed without applying the FWL theorem. There was no inflation in the observed P value distributions.

To evaluate the potential improvement in statistical power, we repeated the empirical analyses included in the current study using the selective inference method and max-$|t|$-test, additionally regressing DNAm values on the confounders. Comparing the P values of the five top CpG sites obtained from the selective inference method and max-$|t|$-test before and after applying the FWL theorem, we found that the P values decreased at all five loci (**Web Figure 4**). The P value at *cg*06430102 exceeded the estimated 450K array-wide threshold after the additional adjustment [17], suggesting that the approach improved statistical power.

Given the evidence observed here, we recommend applying the FWL theorem and regress both the exposure variables and the outcome on confounders before subsequent SLCMA analyses. This approach may effectively increase statistical power and overcome bias due to confounding.

4 Web Tables 1 to 4

Web Table 1: Comparison of effect estimates and confidence intervals of the top CpG sites in the empirical analyses, calculated using the covariance test, selective inference, and max-$|t|$-test.

CpG	First hypothesis chosen	DNAm in unexposed group	DNAm in exposed group	Increase in R^2	Effect estimate	Post-selection inference method	*P* value	Lower 95% CI	Upper 95% CI
cg01370449	Very early childhood (2.5 years of age)	0.2439	0.3341	0.0297	0.084	Max-$\|t\|$-test	1.23E-05	0.0532	0.118
						Covariance test	8.87E-08	0.0501	0.1179
						Selective inference	8.09E-06	0.0493	0.1183
cg06430102	Very early childhood (2.5 years of age)	0.9257	0.8619	0.0368	-0.058	Max-$\|t\|$-test	5.58E-07	-0.0789	-0.0384
						Covariance test	1.69E-09	-0.0789	-0.037
						Selective inference	5.32E-07	-0.0791	-0.0367
cg19170021	Early childhood (4.75 years of age)	0.7342	0.8275	0.0275	0.0958	Max-$\|t\|$-test	5.79E-05	0.0578	0.1374
						Covariance test	6.41E-08	0.0542	0.1373
						Selective inference	1.47E-05	0.0536	0.1378
cg05072819	Early childhood (5.75 years of age)	0.0401	0.0534	0.0305	0.0141	Max-$\|t\|$-test	8.87E-06	0.0089	0.0198
						Covariance test	3.49E-08	0.0084	0.0198
						Selective inference	5.70E-06	0.0083	0.0199
cg05936516	Middle childhood (6.75 years of age)	0.1279	0.1532	0.0311	0.0255	Max-$\|t\|$-test	3.26E-06	0.0164	0.0354
						Covariance test	7.47E-08	0.0156	0.0354
						selective inference	5.43E-06	0.0153	0.0355

Web Table 2: Overlap in most strongly associated loci based on results obtained from the covariance test and the two recommended methods (max-$|t|$-test and selective inference) in the empirical analyses.

Number of top loci	**Selective inference**	**Max-$\|t\|$-test**
10	100%	50%
50	84 %	54%
100	89%	56%
1000	91%	55%
2000	93%	56%
5000	94%	58%

For example, the first line indicates that for the first 10 loci identified by the covariance test, all of them were also among the top 10 based on the selective inference results. However, only half of them appeared among the top 10 identified using the max-t-test.

Web Table 3: Summary of the most popular statistical inference methods used in the SLCMA to identify the best fitting theoretical model.

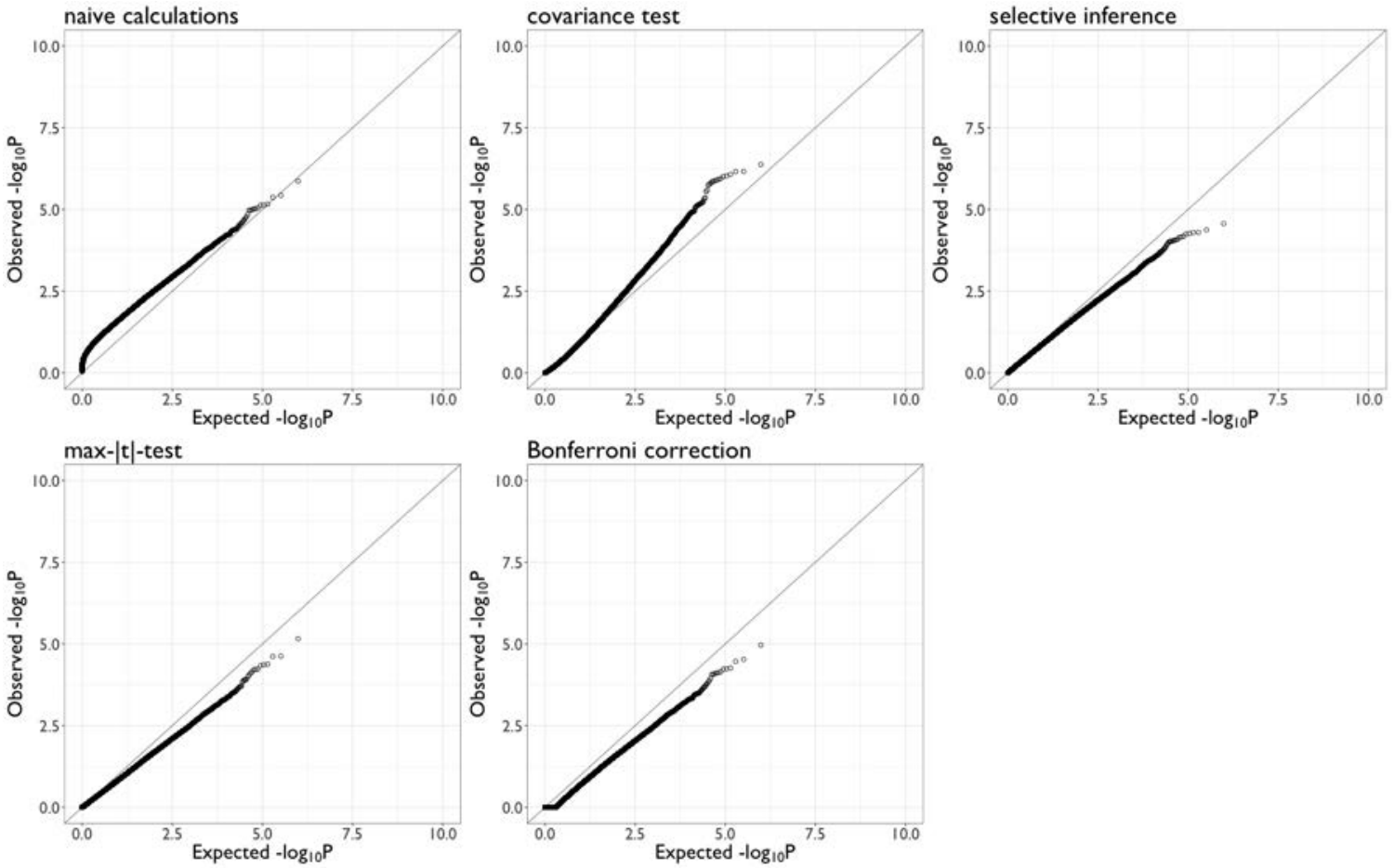
Method	**Model selection procedure**	**Post Selection Inference**		
		Test statistic	**Strategy to address multiple testing and selection**	**Procedure to derive confidence intervals**
Naïve calculations	Forward stepwise regression and least angle regression are equivalent when considering just the predictor with the largest correlation with the outcome	$\hat{\beta}_{OLS} = x_1^T y$	NA	Ordinary least squares (OLS)
Bonferroni correction		$\hat{\beta}_{OLS} = x_1^T y$	Bonferroni correction	NA
Max-$\|t\|$-test		$r_1 = x_1^T y$ where x_1 is the predictor that has the largest correlation with the outcome	Condition the test statistic distribution on it having the maximal correlation with the outcome	Linear transformation of non-cuboid space; can be calculated using existing software
Covariance test		$\lambda_1(\lambda_1 - \lambda_2)/\sigma^2$, where λ_1 and λ_2 are the values of the smoothing parameters at the first and second step of LARS	Condition the test statistic distribution on it having the maximal correlation with the outcome	A modification of the OLS confidence intervals using the corresponding covariance test pvalues
Selective inference		p-value can be shown to be: $\frac{1-\Phi(\lambda_1/\sigma)}{1-\Phi(\lambda_2/\sigma)}$ where Φ is the cumulative distribution function for a standard normal distribution	Conceptualized the selection as responses being in a polyhedral set	Inverting the test statistic

Web Table 4: Estimated family-wise error rate and corresponding 95% CI in a theoretical scenario, after 2 000 simulation experiments

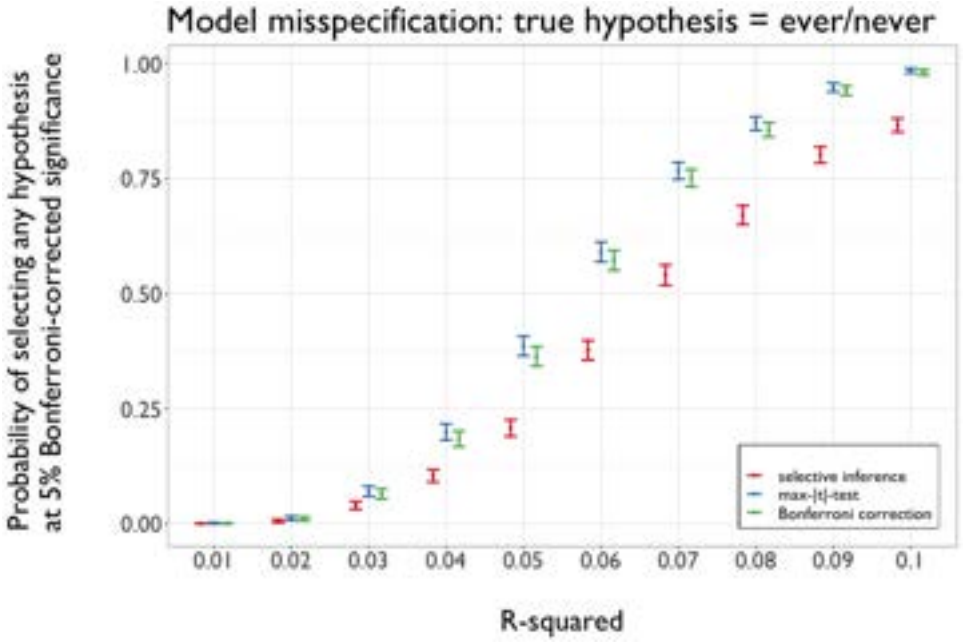
Number of tests	**1**	**10**	**100**	**1000**
Naïve calculation	38.6% (36.5-40.8)	38.8% (36.7-40.9)	38.8% (36.7-40.9)	38.1% (36.0-40.3)
Bonferroni correction	4.9%(3.9-5.8)	4.2% (3.4-5.1)	5.3% (4.4-6.3)	5.1% (4.1-6.0)
Covariance test	6.0% (5.0-7.0)	12.1% (10.7-13.5)	22.5% (20.7-24.3)	42.7% (40.5-44.9)
Selective inference	5.1% (4.1-6.0)	5.2% (4.3-6.2)	5.1% (4.1-6.1)	5.1% (4.1-6.0)
Max-$\lvert t \rvert$-test	5.0% (4.0-5.9)	4.2% (3.4-5.1)	5.3% (4.4-6.3)	5.1% (4.1-6.0)

5 Web Figures 1 to 5

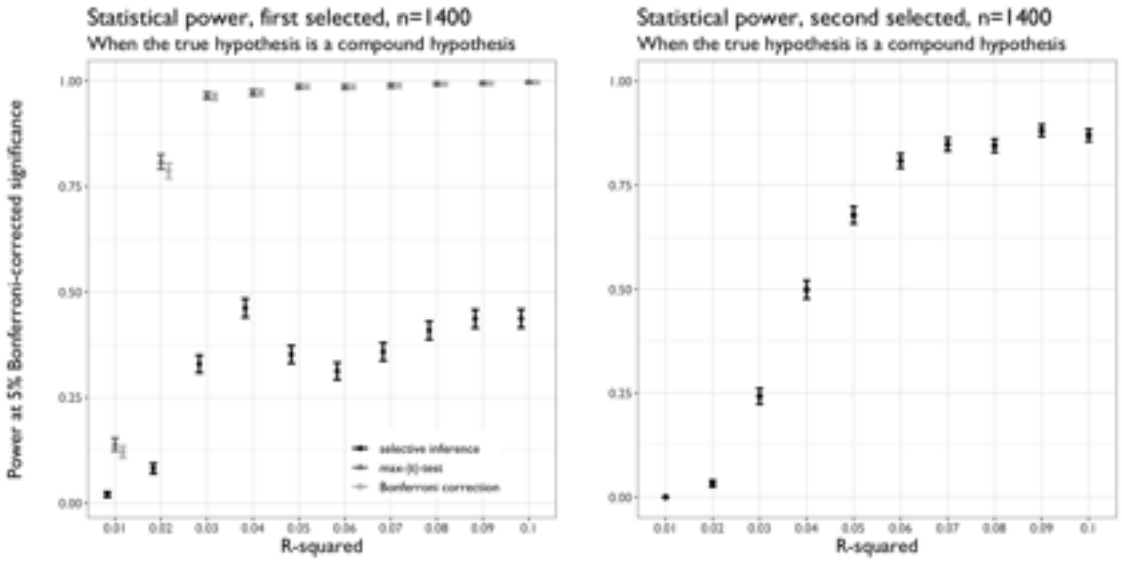
Web Figure 1: Q-Q plots comparing the expected and observed P values simulated under the null for all five methods with empirical outcomes (N=700), where the outcome variables were resampled from observed DNAm values and transformed to M-values.



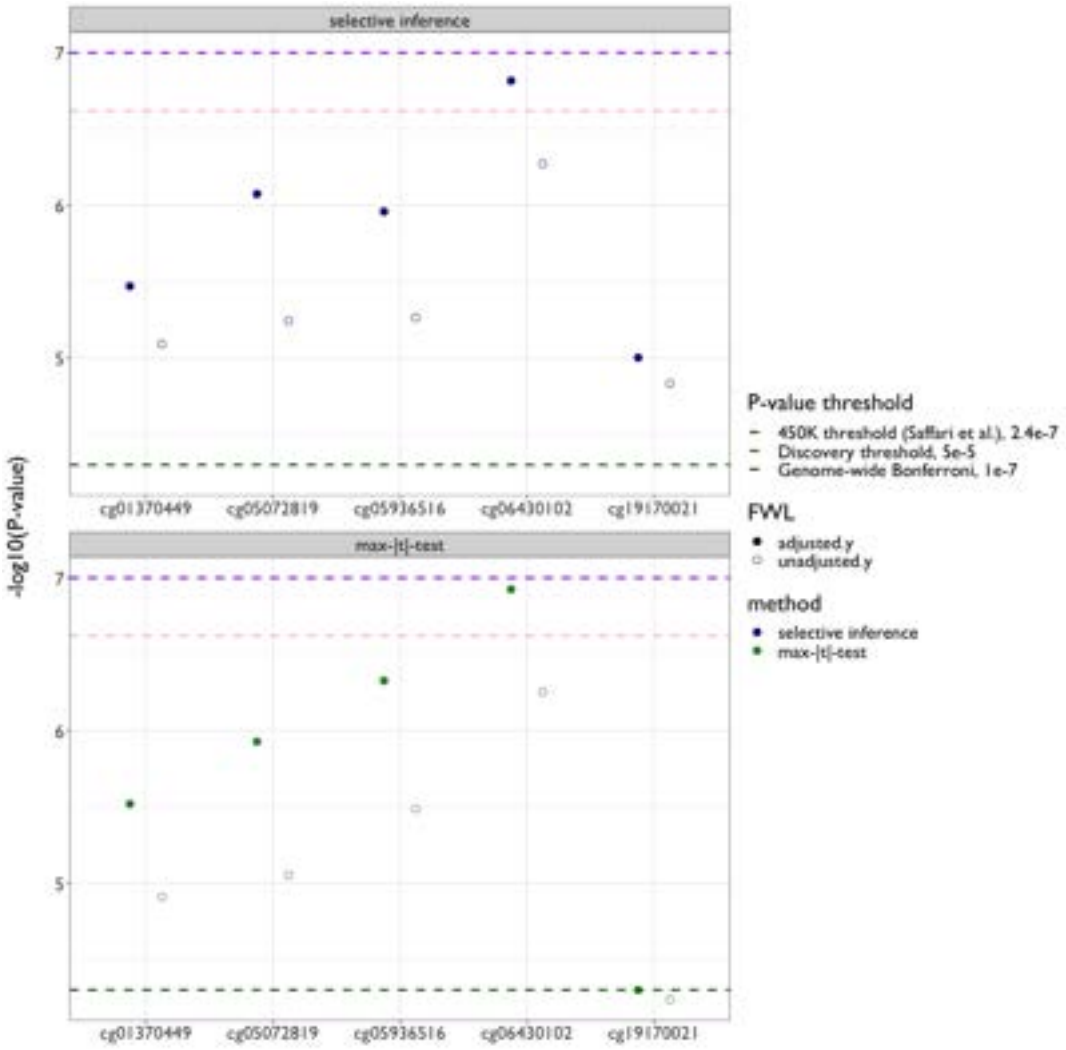
Web Figure 2: Estimated probability of selecting any hypothesis with a 5% Bonferroni corrected P value threshold under model misspecification.



Web Figure 3: Estimated statistical power and corresponding 95% CI in simulated epigenome-wide analyses with increased sample size (n=1400), with varying effect sizes, when the true causal relationship was represented by two hypotheses working in combination.

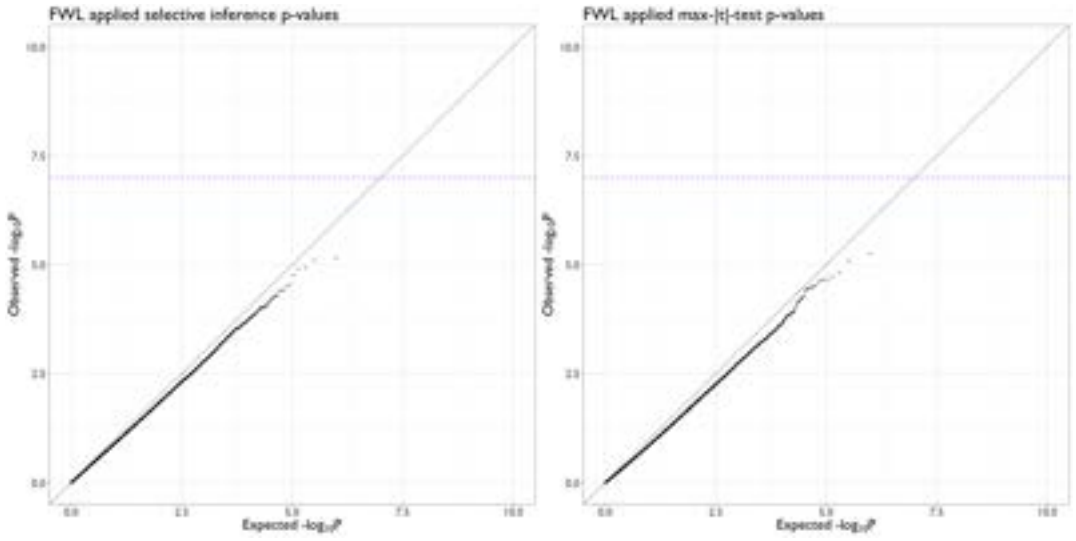


Web Figure 4: Differences in P values of the top CpG sites, before and after applying the FWL theorem, obtained from the selective inference method and max-$|t|$-test



The plot shows the change in $-\log_{10}(p)$ obtained using the two recommended methods: selective inference and the max-$|t|$-test. Open dots represent P values before additionally adjusting for the covariates following the FWL theorem; solid dots represent P values after the FWL adjustment (i.e., regressing the outcome on the covariates and analyzing the residuals). The three dashed lines in different colors denote three commonly used threshold considered in genome-wide DNA methylation studies.

Web Figure 5: Q-Q plots comparing the expected versus observed P values simulated under the null for selective inference and max-—t—-test with empirical outcomes (N=700), after applying the Frisch-Waugh-Lovell theorem to adjust for covariates.



References

[1] Human Tissue Act 2004. http://www.legislation.gov.uk/ukpga/2004/30/pdfs/ukpga_20040030_en.pdf, 2004.

[2] Boyd, A., Golding, J., Macleod, J., Lawlor, D. A., Fraser, A., Henderson, J., Molloy, L., Ness, A., Ring, S., and Davey Smith, G. Cohort Profile: The 'Children of the 90s'—the index offspring of the Avon Longitudinal Study of Parents and Children. *International Journal of Epidemiology 42*, 1 (Feb. 2013), 111–127.

[3] Bühlmann, P., Meier, L., and van de Geer, S. Discussion: "A significance test for the lasso". *Annals of Statistics 42*, 2, 469–477.

[4] Buja, A., and Brown, L. Discussion: "A significance test for the lasso". *Annals of Statistics 42*, 2 (Apr. 2014), 509–517.

[5] Du, P., Zhang, X., Huang, C.-C., Jafari, N., Kibbe, W. A., Hou, L., and Lin, S. M. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics 11* (Nov. 2010), 587.

[6] Dunn, E. C., Soare, T. W., Zhu, Y., Simpkin, A. J., Suderman, M. J., Klengel, T., Smith, A. D. A. C., Ressler, K. J., and Relton, C. L. Sensitive Periods for the Effect of Childhood Adversity on DNA Methylation: Results From a Prospective, Longitudinal Study. *Biological Psychiatry 85*, 10 (May 2019), 838–849.

[7] Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. Least angle regression. *The Annals of statistics 32*, 2 (2004), 407–499.

[8] Fan, J., and Ke, Z. T. Discussion: "A significance test for the lasso". *Annals of Statistics 42*, 2, 483–492.

[9] Fraser, A., Macdonald-Wallis, C., Tilling, K., Boyd, A., Golding, J., Davey Smith, G., Henderson, J., Macleod, J., Molloy, L., Ness, A., Ring, S., Nelson, S. M., and Lawlor, D. A. Cohort Profile: The Avon Longitudinal Study of Parents and Children: ALSPAC mothers cohort. *International Journal of Epidemiology 42*, 1 (Feb. 2013), 97–110.

[10] Frisch, R., and Waugh, F. V. Partial Time Regressions as Compared with Individual Trends. *Econometrica 1*, 4 (Oct. 1933), 387.

[11] Lockhart, R., Taylor, J., Tibshirani, R., and Tibshirani, R. covTest: Computes covariance test for adaptive linear modelling, R package version 1.02.

[12] Lockhart, R., Taylor, J., Tibshirani, R. J., and Tibshirani, R. A significance test for the lasso. *Annals of Statistics 42*, 2 (Apr. 2014), 413–468.

[13] Lovell, M. C. Seasonal Adjustment of Economic Time Series and Multiple Regression Analysis. *Journal of the American Statistical Association 58*, 304 (Dec. 1963), 993.

[14] Lovell, M. C. A Simple Proof of the FWL Theorem. *The Journal of Economic Education 39*, 1 (Jan. 2008), 88–91.

[15] Reid, S., Tibshirani, R., and Friedman, J. A study of error variance estimation in Lasso regression. *Statistica Sinica* (2016).

[16] Relton, C. L., Gaunt, T., McArdle, W., Ho, K., Duggirala, A., Shihab, H., Woodward, G., Lyttleton, O., Evans, D. M., Reik, W., Paul, Y.-L., Ficz, G., Ozanne, S. E., Wipat, A., Flanagan, K., Lister, A., Heijmans, B. T., Ring, S. M., and Davey Smith, G. Data Resource Profile: Accessible Resource for Integrated Epigenomic Studies (ARIES). *International journal of epidemiology 44*, 4 (Aug. 2015), 1181–1190.

[17] Saffari, A., Silver, M. J., Zavattari, P., Moi, L., Columbano, A., Meaburn, E. L., and Dudbridge, F. Estimation of a significance threshold for epigenome-wide association studies. *Genetic Epidemiology 42*, 1 (Feb. 2018), 20–33.

[18] Smith, A. D. A. C., Heron, J., Mishra, G., Gilthorpe, M. S., Ben-Shlomo, Y., and Tilling, K. Model Selection of the Effect of Binary Exposures over the Life Course. *Epidemiology (Cambridge, Mass.) 26*, 5 (Sept. 2015), 719–726.

[19] Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological) 58*, 1 (1996), 267–288.

[20] Tibshirani, R., Tibshirani, R., Taylor, J., Loftus, J., and Reid, S. selectiveInference: Tools for post-selection inference, R package version 1.2.0.

[21] Tibshirani, R. J., Taylor, J., Lockhart, R., and Tibshirani, R. Exact Post-Selection Inference for Sequential Regression Procedures. *Journal of the American Statistical Association 111*, 514 (Apr. 2016), 600–620.

[22] Yamada, H. The Frisch–Waugh–Lovell theorem for the lasso and the ridge regression. *Communications in Statistics - Theory and Methods 46*, 21 (Nov. 2017), 10897–10902.