ORIGINAL ARTICLE

Performance of variable selection methods for assessing the health effects of correlated exposures in case–control studies

Virissa Lenters,[1] Roel Vermeulen,[1,2] Lützen Portengen[1]

► Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/oemed-2016-104231).

[1]Division of Environmental Epidemiology, Institute for Risk Assessment Sciences, Utrecht University, Utrecht, The Netherlands
[2]Departmentof Epidemiology, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, The Netherlands

Correspondence to
Dr Lützen Portengen, Institute for Risk Assessment Sciences, Utrecht University, PO Box 80.178, 3508 TD Utrecht, The Netherlands; L.Portengen@uu.nl

RV and LP contributed equally.

Received 29 November 2016
Revised 16 August 2017
Accepted 22 August 2017
Published Online First
25 September 2017

ABSTRACT

Objectives There is growing recognition that simultaneously assessing multiple exposures may reduce false positive discoveries and improve epidemiological effect estimates. We evaluated the performance of statistical methods for identifying exposure–outcome associations across various data structures typical of environmental and occupational epidemiology analyses.

Methods We simulated a case–control study, generating 100 data sets for each of 270 different simulation scenarios; varying the number of exposure variables, the correlation between exposures, sample size, the number of effective exposures and the magnitude of effect estimates. We compared conventional analytical approaches, that is, univariable (with and without multiplicity adjustment), multivariable and stepwise logistic regression, with variable selection methods: sparse partial least squares discriminant analysis, boosting, and frequentist and Bayesian penalised regression approaches.

Results The variable selection methods consistently yielded more precise effect estimates and generally improved selection accuracy compared with conventional logistic regression methods, especially for scenarios with higher correlation levels. Penalised lasso and elastic net regression both seemed to perform particularly well, specifically when statistical inference based on a balanced weighting of high sensitivity and a low proportion of false discoveries is sought.

Conclusions In this extensive simulation study with multicollinear data, we found that most variable selection methods consistently outperformed conventional approaches, and demonstrated how performance is influenced by the structure of the data and underlying model.

What this paper adds

- It is well established that assessing associations between multiple exposures and a binary health outcome using standard logistic regression models can result in false positive and imprecise effect estimates.
- There is limited empirical support for the performance of alternative statistical approaches; particularly for low-dimensional scenarios (fewer exposures than study participants) with moderately to highly correlated exposures.
- This simulation study demonstrates how nine statistical approaches perform under a wide range of scenarios.
- Penalised regression models, which shrink regression coefficients towards or to zero, performed well and offer an attractive approach.

To better estimate the modifiable components of disease risk, environmental and occupational epidemiologists are increasing their efforts to capture the myriad of exposures humans come in contact with. Researchers are using a variety of tools to perform multifaceted exposure assessment in observational epidemiology studies, including using data from questionnaires, sensors, geospatial modelling, job exposure matrices, biomonitoring and high-throughput molecular analyses. As researchers gather increasingly richer data sets, they may find themselves confronted with complex exposure profiles that render identification of exposure–outcome associations using conventional statistical analyses challenging.

Studies often assess one exposure, or sometimes one chemical class or closely related group of exposures, in relation to a health outcome. A shift from single-exposure modelling to multiple-exposure (multipollutant or chemical mixture) modelling has been advocated to better identify and estimate the independent effects of exposures.[1] Selection and estimation using conventional methods for multipollutant modelling, such as stepwise regression and multiple regression, may suffer from locally rather than globally optimal models, influenced by the order variables enter the model, or may yield unstable estimates if exposures are multicollinear. Use of more advanced statistical methods has gained traction in the past few years, particularly in chemical contaminant and air pollution epidemiology.[2–5] Several (families of) variable selection methods have more recently been developed (see Chadeau-Hyam *et al*[6] for a review). However, there are limited data on the relative efficacy of these methods for analysis of data generally relevant for occupational and environmental epidemiologists; that is, with a limited number of observations, relatively few true associations, complex correlation structures and modest effect sizes.

Recently published simulation analyses have evaluated methods for the analysis of continuous

► http://dx.doi.org/10.1136/oemed-2017-104807

Check for updates

To cite: Lenters V, Vermeulen R, Portengen L. *Occup Environ Med* 2018;**75**:522–529.

Table 1 Methods applied for analysis of multiple exposures and a binary health outcome

Method	R package: function(s)	Implementation
Univariable logistic regression	stats: glm	Selection: p<0.05
Univariable-FDR logistic regression[9]	stats: glm, p.adjust	Selection: FDR<0.05
Multivariable logistic regression	stats: glm	Selection: p<0.05
Stepwise logistic regression	stats: step, glm	Selection: smallest Akaike information criterion
sPLS-DA[12 13]	spls: cv.splsda, splsda	Model (K, η) tuned via CV. Selection: $\beta \neq 0$
Lasso regression[14 42]	glmnet: cv.glmnet, glmnet	Model (λ) tuned via CV. Selection: $\beta \neq 0$
Elastic net regression[15 42]	glmnet: cv.glmnet, glmnet	Model (α, λ) tuned via CV. Selection: $\beta \neq 0$
Bayesian lasso regression[17]	reglogit: reglogit	500 burn-in plus 1000 iterations. Selection: If 95% highest posterior density interval did not include 0.
Boosted regression[19 43]	mboost: cvrisk, glmboost	≤1500 iterations; CV to determine the stopping iteration. Selection: $\beta \neq 0$

CV, cross-validation; FDR, false discovery rate; sPLS-DA, sparse partial least squares discriminant analysis.

outcomes considering data structures inspired by the pregnancy exposome[7] and air pollution epidemiology.[8] We extended this work by (1) studying a binary outcome, as much of epidemiological research deals with data from case–control studies and presence/absence of disease, (2) focusing on a larger set of simulation scenarios, and (3) by exploring in what way different characteristics of the simulated exposure matrix and the strength of the exposure–outcome association affect the performance of variable selection methods.

We evaluated dimension reduction, penalisation and boosting approaches which are readily implemented in standard software, and pursued the scenario in which there are no prior hypotheses regarding candidate associations, or where this information is ignored; so following a data-driven approach to variable selection.

METHODS

Variable selection methods

We briefly describe each method used to estimate associations between simulated exposures and a binary health outcome. Additional details are provided in table 1 and the online supplementary appendix 1. We used R software V.3.2.1 (R Foundation for Statistical Computing, Vienna, Austria) to simulate and analyse our data.

Univariable(-FDR) and multivariable. The most frequently used strategy for binary classification in epidemiology is to assess separate single-exposure logistic regression models (possibly adjusted for a limited set of known or suspected confounders), referred to as univariable (or univariate) regression. Another strategy is multivariable (or multiple) regression, in which all exposure variables are included in one logistic regression model. We fit generalised linear models with the logit link function and maximum likelihood estimation. Selection of exposures was based on p values, with a value of 0.05 considered as the cut-off. We also evaluated univariable regression with selection based on controlling the false discovery rate (FDR <0.05, using the Benjamini-Hochberg method[9]). This approach is increasingly being applied to adjust for multiplicity, and has been applied in the so-called environment-wide association study approach.[10]

Stepwise. We implemented forward-backward stepwise logistic regression in which one variable is successively added or deleted at each step, based on minimising the Akaike information criterion. This is a widely used automated selection procedure.

sPLS-DA. Partial least squares (PLS) regression[11] is a dimension reduction method which projects the outcome and exposure variables onto a smaller number of orthogonal latent variables, called components, maximising their covariance. Sparse PLS (sPLS) regression[12 13] introduces variable selection by indirectly imposing an L_1 (lasso type) penalty on the direction (loading) vector defining the projection, so that only a subset of the most informative exposure variables contributes to the components. sPLS can be used for classification problems by inputting the outcome as a dummy matrix in the PLS algorithm, and using the resulting component scores in a linear discriminant analysis (DA). Model complexity is determined by the number of components (K) and the level of penalisation (η), which were tuned using two-dimensional cross-validation. We selected the model which corresponded to the minimum misclassification error rate in the primary analyses, and modified the default cross-validation within the R package to include an intercept-only model.

Lasso. Penalisation approaches, also referred to as regularisation or shrinkage methods, shrink coefficients towards zero, with the amount of shrinkage inversely proportional to the contribution of the variable to the overall model fit (ie, important variables are shrunken less). This shrinkage introduces bias but also decreases the variance and generally leads to more stable models, especially for multicollinear data. The most popular penalty is the lasso (least absolute shrinkage and selection operator) penalty, which is proportional to the sum of the absolute values of the coefficients.[14] Variable selection (sparsity) is achieved because coefficients for a subset of variables may be shrunk exactly to zero.

Elastic net. Elastic net regression is similar to lasso regression, but uses a weighted sum of lasso and ridge regression penalties.[15] The ridge regression penalty is proportional to the sum of the squared regression coefficients, which results in shrinkage of the coefficients towards zero, but not to zero exactly, and for coefficients for highly correlated variables towards a common value. Whereas lasso regression tends to select only one variable from a group of correlated variables, elastic net can coselect a group of correlated variables (due to the ridge penalty), while still performing variable selection (due to the lasso penalty). A second tuning parameter (α) controls the balance between the lasso and ridge penalties; we optimised α and the amount of penalisation (λ) using two-dimensional cross-validation. For both lasso and elastic net regression, we selected the most parsimonious model which corresponded to the minimum plus one SE in binomial deviance in the primary analyses—a more stringent optimisation criterion than the minimum, and one that has been advocated for variable selection.[16]

Bayesian lasso. Variable selection in Bayesian models can be achieved using independent Laplace (or double exponential) prior distributions centred at zero to estimate regression coefficients, also referred to as regularised logistic regression.[17 18] A hyperprior controls the shape of these Laplace priors, where the optimal amount of shrinkage is adaptively determined from the

Table 2 Parameters tested in simulation scenarios*

n†	p‡	ρ§	q¶	OR**
200 (9000)	10 (9000)	0.0 (9000)	0 (2700)	1.0 (2700)
500 (9000)	20 (9000)	0.4 (9000)	1 (8100)	1.2 (8100)
1000 (9000)	50 (9000)	0.8 (9000)	2 (8100)	1.5 (8100)
			5 (8100)	2.0 (8100)

*Parameter levels and (the number of simulated data sets generated per level). 100 simulations were generated for each of the 270 [=3×3×3×(1+3×3)] simulation scenarios, where effect size 1.0 only applies to simulations with 0 effective variable.
†The sample sizes.
‡The number of exposure or predictor variables.
§The correlation levels in a square diagonal matrix whose diagonal elements are 1.0 with uniform correlations between exposures (boxed correlations were simulated in a sensitivity analysis).
¶The number of effective or truly associated exposures.
**The effect sizes.

data itself, achieving approximate multiplicity control. A variable was considered to be selected if its 95% highest posterior density interval did not include 0.

Boosted. We fitted generalised linear models using a component-wise gradient boosting algorithm,[19 20] which is an iterative technique that aims to optimise predictive accuracy by combining a set of weak classifiers. With each iteration step, the variable that results in the greatest reduction in error based on the negative binomial log-likelihood loss function is selected. A small proportion (here 10%) of the fit for the selected variable is added to the current estimate, and the residuals are then used in subsequent iterations. Estimates for the variable selected at each iteration are aggregated. Variable selection is achieved by restricting the number of boosting iterations, and cross-validation with the minimum binomial deviance criterion was used to determine the optimal number of iterations.

Simulation study

We designed our simulation study to be representative of data structures that are common in environmental and occupational epidemiology. We performed a Monte Carlo simulation, varying five aspects of the data (listed in table 2), resulting in a total of 270 different scenarios. For each scenario we simulated 100 data sets, yielding a total of 27 000 different data sets.

First, the matrix of continuous exposures X was simulated by sampling 100 000 observations from a multivariate normal distribution with a mean of zero, SD of 1, and covariance matrix with off-diagonal elements with a uniform correlation level of ρ. X was column mean centred and standardised to unit variance. Second, the case/control status for the outcome vector Y was simulated from a Bernoulli distribution, with probability $\Pr(Y=1|X)=1/\{1+\exp[-(\beta_0+\beta_X X)]\}$. Parameter β_0 was chosen as to be sufficiently large to ensure that the simulation resulted in enough cases and controls for the scenario-specific study size, while β_X was fixed according to the magnitude of the desired effect estimate, the $\mathrm{OR}=\exp(\beta_X)$. An equal number of cases and controls were sampled without replacement and used to construct the final data set for the analyses.

Evaluation criteria

To assess the relative classification performance of the variable selection methods, we tallied the number of true positive, false positive, true negative and false negative associations (see online supplementary table S1) across simulations, and computed:

1. The sensitivity: the proportion of true associations that were correctly identified, also referred to as recall. No value for sensitivity was computed when the number of true associations equalled zero.
2. The false discovery proportion (FDP): the proportion of positive findings (associations identified as significant) that were incorrectly identified. FDP was defined as zero when the number of positive findings was zero. The complement of FDP (1–FDP) is equivalent to the positive predictive value, also referred to as precision.

The simulation design is imbalanced in that the number of true negatives exceeds the number of true positives. As such, we compare 1–FDP to the sensitivity, akin to a precision-recall curve, rather than comparing the sensitivity to the false positive rate (the proportion of true null associations identified incorrectly as significant), as is shown in receiver operating characteristic curves commonly used to assess the performance of classifiers.[21] We also consider various weightings of the F measure, the weighted harmonic mean of sensitivity and 1–FDP, where either metric can be assigned greater importance (differential weighting) or equal importance (balanced weighting).[22]

3. The mean squared error (MSE): to assess the estimation performance, we computed a combination of the (squared) bias of the estimated coefficients plus its variance across s simulations, $\frac{1}{s}\sum\left(\hat{\beta}-\beta\right)^2$.

We evaluated the overall performance of the methods, across all simulation scenarios, and across all parameter levels. We focus on the results obtained upon varying the three parameters that may be known or assessed in advance by the researcher (ie, the sample size, the number of exposure variables and the correlations between exposure variables), as opposed to the parameters that are estimated but remain uncertain in most real-life analyses (ie, the number of true associations and the effect sizes).

As a sensitivity analysis we repeated all 27 000 simulations using a blocked correlation structure (instead of a single, uniform correlation structure), where we subdivided the correlation matrix into two equally sized submatrices that had the same uniform correlation between exposures within the *same* submatrix ($\rho=0.4$ or $\rho=0.8$), but where there was no correlation between exposures in *different* submatrices ($\rho=0.0$). Additionally, we retested sPLS-DA and frequentist penalisation models with alternative optimisation criteria; the minimum plus one SE and the minimum cross-validation error, respectively.

RESULTS

Overall performance

We used nine different selection strategies to analyse the 27 000 simulated data sets. The overall results, averaged across all simulation scenarios, are presented in figure 1A. A higher sensitivity and lower FDP indicate superior performance in selecting true associations and fewer false associations, thus if a researcher considers sensitivity to be equally as important as FDP, then methods located in the upper right quadrant of the plot (ie, 1–FDP-sensitivity or precision-recall space) are more favourable. The area of the bubbles is proportional to the median MSE of the coefficient estimates, and smaller bubbles therefore reflect more stable and precise estimation of the true magnitude of associations. The lasso and elastic net regression methods appear to be the best performing methods overall for achieving a sensitivity:FDP balanced selection performance, with the elastic net achieving slightly higher sensitivity and lasso regression achieving a somewhat lower FDP, and both yielding nearly the lowest MSE values. When we examine these three performance criteria separately (online supplementary figure S1), we note that

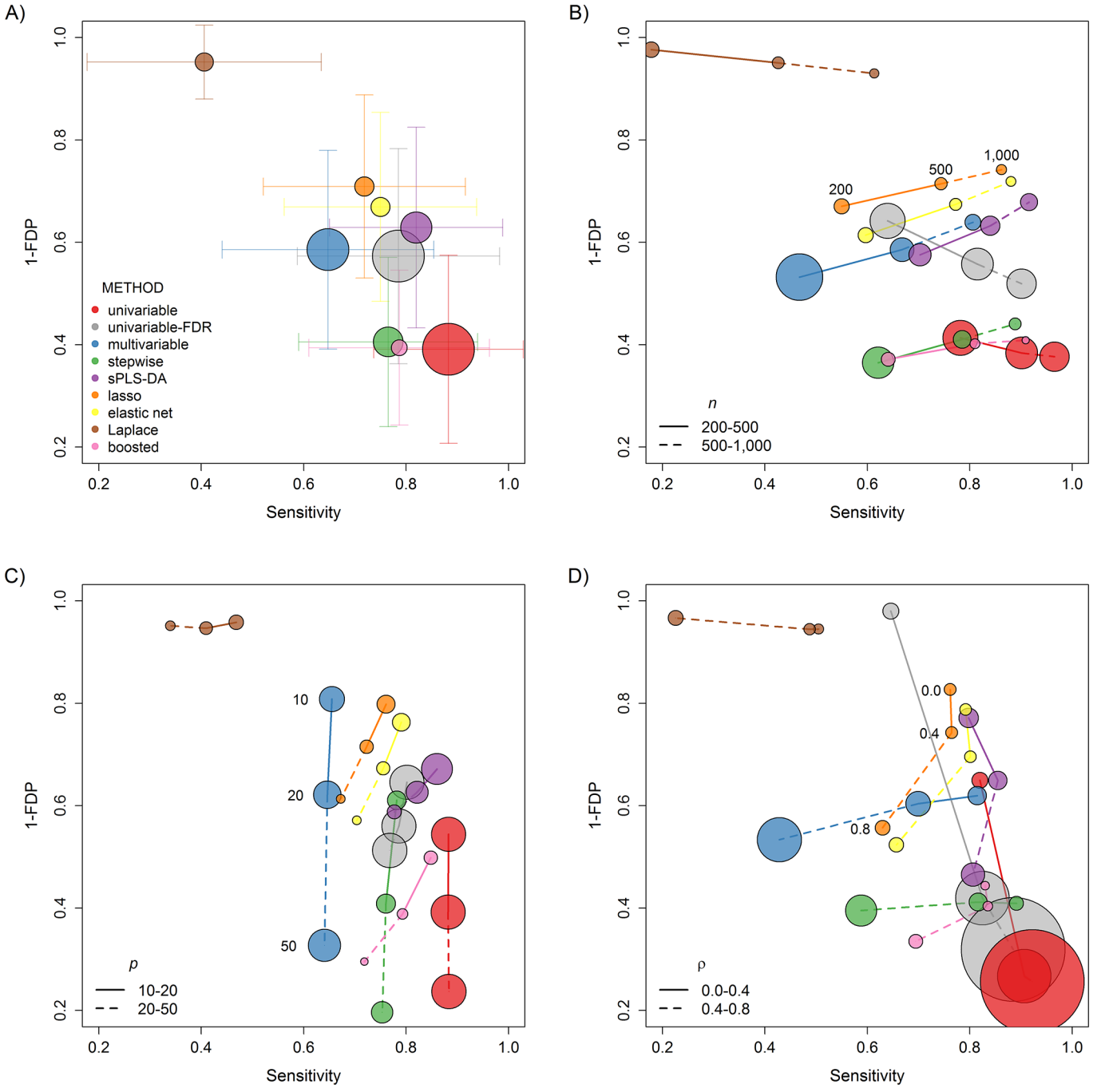

Figure 1 Performance of the variable selection methods (A) across all 270 simulation scenarios and stratified by (B) sample size, (C) number of exposure variables and (D) correlation levels. The mean sensitivity is plotted against 1 minus the mean false discovery proportion (1–FDP), where the size of each bubble is proportional to the median mean squared error (MSE). Error bars represent half SD values of sensitivity and 1–FDP. Note that the axes are truncated at 0.2, and the scaling of MSE values is relative and differs between plots. FDR, false discovery rate; sPLS-DA, sparse partial least squares discriminant analysis.

the conventional modelling approach, univariable regression, has both the best sensitivity and worst FDP ranking. Univariable with FDR control yields an improved selection performance, although the estimation performance remains unchanged. A far greater proportion of MSE values are large for univariable regression compared with the other methods (the 75th percentile was 0.31 for univariable and univariable-FDR vs 0.06 for multivariable and 0.01 for elastic net regression).

Researchers may value sensitivity more highly than FDP or vice versa, depending on the focus of research question. If we consider a composite of sensitivity and 1–FDP, the median F measure across simulation scenarios, then lasso and elastic net regression have the highest ranking whether sensitivity and 1–FDP have a balanced weighting (F_1) or whether sensitivity is a quarter or half as important ($F_{0.25}$, $F_{0.5}$) or twice or four times as important (F_2, F_4) as 1–FDP. The ranking of the methods across all simulations at F_1 (elastic net>lasso>sPLS-DA>boosted>stepwise>multivariable>univariable>univariable-FDR>-Bayesian lasso) is fairly consistent across $F_{0.25}$ to F_4 differential weightings (with only boosted, stepwise and multivariable

swapping the middle-ranking positions; online supplementary table S2).

Influence of parameters on performance

We examined the performance across varying levels of the simulation parameters, starting with those that are likely to be known by researchers before the statistical analyses (figure 1B–D). Bayesian lasso consistently occupied the upper left quadrant, displaying very low FDP, but poor overall sensitivity. The other methods clustered between a 1–FDP ranging from around 0.2 to 0.8, and sensitivity around 0.4–0.9. The relative position of the methods in the FDP and sensitivity space generally remained consistent upon varying the parameters. The penalised regression methods (lasso, elastic net and Bayesian lasso) and boosted regression consistently displayed the smallest MSE values across the levels of the parameters (see also online supplementary figure S3). Mean sensitivity improved substantially with increasing sample size for all methods, and most so for Bayesian lasso (+0.44 from n=200 to n=1000) and multivariable regression (+0.34). The FDP was less affected by sample size; most methods showed minor improvements (up to +0.11 in mean 1–FDP), and Bayesian lasso and univariable(-FDR) showed a minimal decline. The median MSE for multivariable regression at the smallest sample size (n=200) was 1.8 times larger than the next largest MSE (univariable(-FDR) at n=200).

Sensitivity was minimally affected by the number of exposure variables (≤–0.13 for an increase from 10 to 50), whereas the 1–FDP was strongly affected for conventional methods (–0.31 to –0.48) and moderately affected for univariable-FDR (–0.13) and most of the variable selection methods (–0.08 to –0.20; except for the Bayesian lasso, which was negligibly affected). This may have been expected since there was no direct multiplicity adjustment in variable selection methods, except for the adaptable prior in Bayesian lasso, and this clearly demonstrates that cross-validation is no substitute.

For increasing correlation levels, the patterns were less consistent across methods. Most methods seemed tolerant to small-to-moderate correlations, with the sensitivity and/or FDP affected more seriously when the correlation levels increase from ρ=0.4 to ρ=0.8, as compared with an increase from ρ=0.0 to ρ=0.4. The exceptions were univariable and univariable-FDR, which showed a large drop in 1–FDP from the uncorrelated to correlated scenarios. With respect to the F measure, elastic net and lasso had the highest $F_{0.25\text{–}4}$ measure rankings upon stratification by correlation level, with the exception of univariable-FDR in the uncorrelated scenarios (which had the highest ranking at ρ=0.0; online supplementary table S2).

We also examined the performance for parameters that may be estimated but remain uncertain (online supplementary figure S2). As the number of effective (truly associated) exposures increased from 1 to 5, there was a moderate increase in the proportion of false discoveries for most methods, except the penalised regression methods. The sensitivity consistently increased across methods as the strength of the association increased.

To gain more insight into the behaviour of the methods, we stratified by all three parameters which can be assessed in advance by the researcher. In figure 2 we compare the performance of more conventional methods, univariable-FDR and stepwise regression, with the less frequently applied elastic net regression (results for each method are presented in online supplementary figure S3). Methods located at the top of the panel, with one panel for each performance metric, indicate superior performance. Elastic net clearly outperforms stepwise with respect to FDP, and also outperforms univariable-FDR in scenarios with a higher number of exposures (p=20 and p=50). All three methods exhibit a similar sensitivity with p=20 exposures, whereas stepwise has the highest sensitivity with p=10 and univariable-FDR with p=50. The MSE values are lowest for elastic net, modestly reduced compared with stepwise and markedly reduced compared with univariable-FDR in scenarios with p=20 and p=50 exposures.

Sensitivity analyses

Upon repeating the simulation analysis using a blocked correlation structure, the selection and estimation performance was similar to the simulations with a uniform correlation structure, with the exception of a marked decrease in MSE for univariable(-FDR) with correlated exposures in a blocked structure (online supplementary figure S4). Optimising the sPLS-DA model based on the minimum plus one SE cross-validation prediction error, rather than the minimum error, led sPLS-DA to have a negligible FDP but poor sensitivity. The changes in performance for lasso and elastic net with a switch to optimisation based on the minimum error were less substantial; a moderate increase in FDP and minor increase in sensitivity were observed (online supplementary figure S5). As a post hoc analysis, we evaluated multivariable regression with FDR control, a relatively infrequently applied approach. It showed a similar performance as Bayesian lasso. There was a pronounced improvement in FDP and a moderate drop in sensitivity compared with the primary multivariable analyses without correction for multiplicity (online supplementary figure S6).

DISCUSSION

In this simulation study, based on a case–control design, modern variable selection approaches nearly always outperformed conventional logistic regression approaches in recovering the underlying causal model when exposures were correlated. Single-exposure models (univariable either with or without FDR control) and stepwise regression showed high FDP and MSE values in many scenarios, partly attributable to lack of control for confounding by correlated exposures for the single-exposure models and overfitting for the stepwise regression models. Boosted regression also had a high FDP, although the average MSE was low. Full multivariable regression modelling yielded high MSE and moderately high FDP, especially with higher correlations among exposures. The Bayesian lasso had a very low FDP, although at the cost of greatly reduced sensitivity. Elastic net and lasso regression, and sPLS-DA performed best on average with respect to identifying associations across the wide array of simulation scenarios, for a balance of sensitivity and low FDP, and also across more lenient exploratory versus more stringent confirmatory statistical inference rankings (ie, $F_{0.25}$ to F_4), although elastic net and lasso regression performed better with respect to estimation performance (ie, lower MSE).

Compared with lasso regression, elastic net regression was slightly more sensitive but also had a slightly higher FDP. The two-dimensional cross-validation often resulted in α close to 1 (the lasso setting), so the general similarity is not surprising. Fixing α to a smaller value, such as 0.5, led to slightly higher sensitivity accompanied by greater increases in FDP for elastic net compared with lasso in scenarios with correlated exposures (data not shown). A more complex correlation structure may also have resulted in more divergent results between elastic net and lasso regression. The MSE values were smaller than for most other methods, which indicates that associations were

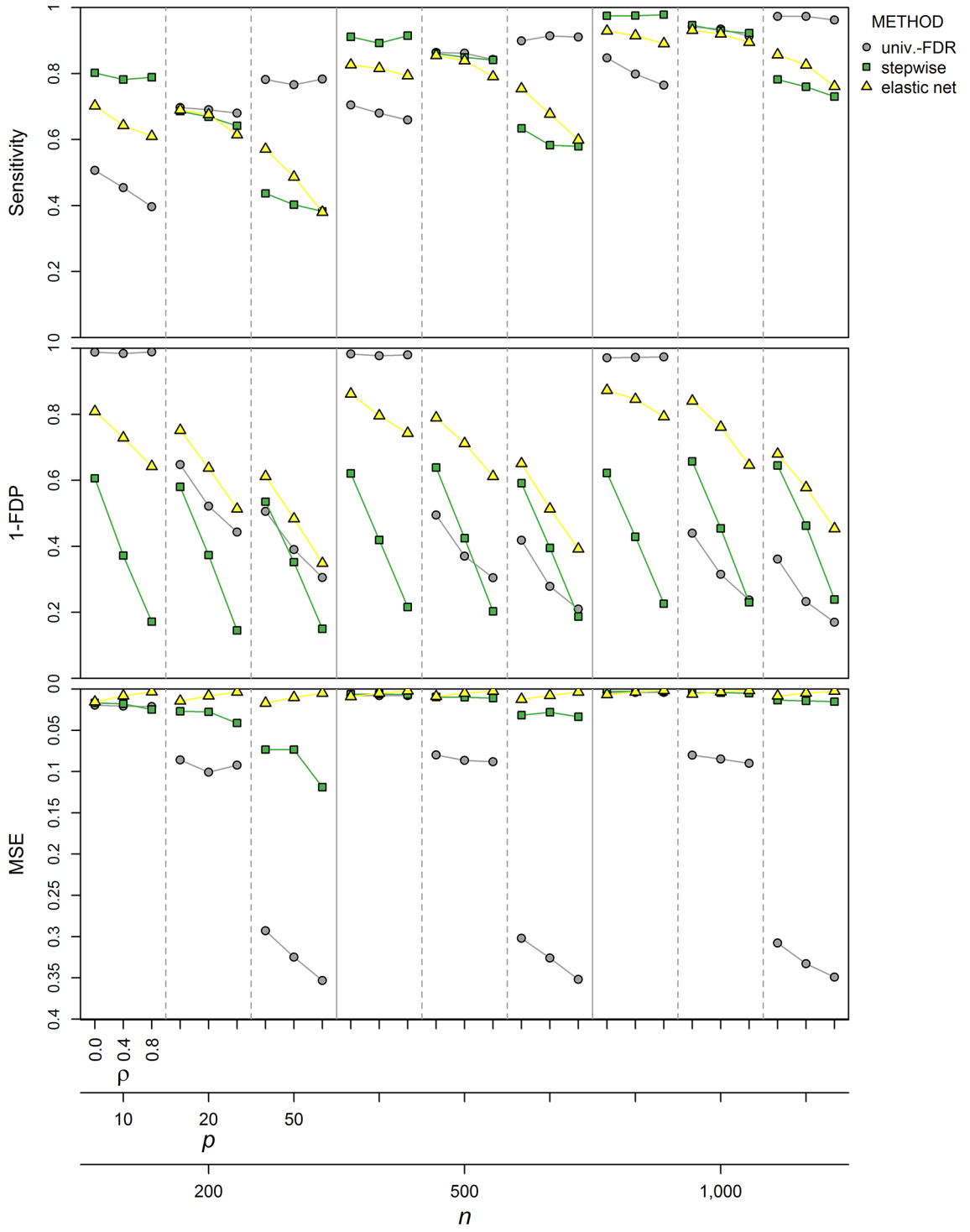

Figure 2 Sensitivity and false discovery proportion (FDP; mean values) and mean squared error (MSE; median values) stratified by the three a priori known simulation scenario parameters for univariable regression with false discovery rate (FDR) correction, stepwise regression and elastic net regression.

estimated more precisely, even though penalisation results in estimates that are biased towards the null. Alternative penalties have been developed which reduce this shrinkage bias for larger coefficients, such as non-convex and adaptive penalties (online supplementary appendix 1).

A drawback of frequentist penalisation methods is that these do not yield valid SEs or other uncertainty measures. Inference based on p values conditional on the selected subset (postselection inference), sequential FDR control and selection based on subsampling stability have recently been explored,[23–25] but are computationally demanding or are not yet available in standard statistical packages and were therefore not included in our simulation study. The Bayesian lasso does provide (samples from) the full posterior for effect estimates, which can be used to draft credible intervals.[26] The Bayesian lasso showed a very low sensitivity in the present simulation study, particularly in scenarios with a smaller sample size (n=200) or a larger number of exposure variables (q=50), reflecting that the prior corrects for multiplicity,[27] in line with multivariable regression with FDR control. Using a slightly lower threshold for variable selection resulted

in behaviour more similar to the frequentist penalised methods (data not shown). In addition to selection criteria, the choice of the sampling (ie, Markov chain Monte Carlo) algorithm and (hyper)prior have been shown to influence performance of Bayesian variable selection approaches.[28 29] Other simulation studies with a higher range of effective variables[30] and high-dimensional scenarios[17] found that Bayesian lasso outperformed frequentist lasso. sPLS-DA performed nearly as well at selecting associations in the present simulation study, although MSE values were somewhat larger than for the penalised regression models. The selection performance of sPLS-DA was much more sensitive to the model optimisation criteria than the penalised regression methods. The coefficient for each selected variable in a sPLS-DA model is calculated by summing over latent components, and this renders the resulting coefficient difficult to interpret unless the latent components have a direct interpretation. Another difficulty with sPLS-DA (and most other dimension reduction methods) is that there is no direct way to adjust for confounding. One option is to first preadjust the outcome and exposure variables for the confounding variables,[31] and use the residuals from these models as input for the sPLS-DA model. (s)PLS-DA may prove more useful for predictive modelling, pattern recognition or delineating clusters than for variable selection.

Boosted regression exhibited an undesirably high FDP. This may be because we relied on minimisation of the cross-validated risk for that method, which is known to result in models that have good predictive ability, but that often include too many variables, especially when the signal-to-noise ratio is low. We demonstrated that several methods (lasso and elastic net and especially sPLS-DA) are sensitive to the model optimisation criteria. This deserves more attention in future simulation studies.[32]

There are many other methods for variable selection (also called model selection), and the battery of available approaches is expected to rapidly evolve; for instance, Bayesian kernel machine regression[33] has recently been proposed as a suitable approach. Selection methods have more frequently been evaluated for high-dimensional data (p>>n), as variable selection methods have proven particularly useful for these data structures, such as genetics and omics data.[34] These simulations have limited generalisability because environmental and occupational exposure sets are often characterised by denser correlations than most genomics data sets.[35] There are also more evaluations of linear regression than of other link functions, as they are generally computationally less demanding and have relatively high power. In a recent simulation based on a low-dimensional pregnancy exposome data set (p=237 exposures and q=0–25 truly associated) and linear modelling, Agier *et al*[7] found that deletion/substitution/addition[36] and a stochastic search algorithm, Graphical Unit Evolutionary Stochastic Search,[37] performed best; we did not include these methods because either a current R package was not available or the method was too computationally demanding for a large simulation study, and the latter method cannot be applied in a logistic regression framework. They also found that elastic net performed reasonably well. In an evaluation of multipollutant linear modelling (p=4–20), Sun *et al*[8] also report that lasso regression performed well. We evaluated a relatively small number of exposures compared with the higher dimensional model space that is expected to be considered in future environment/exposome-wide association studies. Nevertheless, the patterns in performance—notably in the presence of correlated exposures—would apply for the conventional methods and also be expected to apply for the more novel variable selection methods in scenarios with many more exposures.

A limitation of the present study is that simulation results entail some uncertainty; however, increasing the number of simulations (presently 100 per scenario aggregated over various combinations of scenarios) would not be expected to alter the trends or overall inferences. Not unexpectedly, we found that performance was highly dependent on most of the parameters we included in the design of the simulations. We observed that the sample size and strength of the association had the greatest influence on the sensitivity of methods, and selection performance of modern variable selection methods was generally more robust to changes in correlation levels. We tested parameters that we hypothesised would be most influential. Future work could assess others that are highly relevant in environmental and occupational epidemiology, such as the effect of measurement error (due to, eg, limitations in the precision of analytical assays or error in exposure assessment models), skewed exposure distributions, interactions or effect modification by other exposure variables, non-linear associations, mixed models and more complex correlation structures. It has been shown that measurement error, and correlated measurement error, usually attenuate effect estimates in two-exposure modelling of chemical biomarkers and air pollutants.[38 39] Furthermore, model mis-specification, for instance due to inclusion of a variable that improves prediction irrespective of the causal structure, can induce bias due to omitted variable bias and collider stratification bias. Although the magnitude of this bias amplification might often be expected to be relatively minor compared with the reduction in residual confounding bias from correlated coexposures, the cumulative bias is a complex function of many factors including covariance and the strength of each association, and this is an area of methodological development that warrants further attention in variable selection modelling.[40 41]

Simultaneously assessing multiple exposures in an epidemiological analysis, as in the recently championed environment-wide association study approach, reduces selective reporting and publication bias. Variable selection approaches, which mitigate the problem of multicollinearity, can complement or perhaps replace the single-exposure approach to environment-wide association studies. This study offers some guidance for choosing among various modelling approaches considering the structure of the data and whether the aim of the analysis is exploratory with a preference for sensitivity or confirmatory with a preference for specificity. While no single method was infallible in the presence of correlations, we demonstrated that several variable selection approaches yielded improved effect estimation and selection accuracy, and that penalised regression approaches in particular were an attractive option across an array of analytical scenarios.

Contributors LP and RV conceived and designed the analysis. LP performed the simulations and statistical analyses, with contributions from VL. All authors were actively involved in interpretation of results. VL drafted and revised the manuscript, with contributions from LP and RV. All authors approved the final version of this manuscript.

Competing interests None declared.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement The R code for the simulation and analyses is available upon request.

REFERENCES

1 Dominici F, Peng RD, Barr CD, *et al*. Protecting human health from air pollution: shifting from a single-pollutant to a multipollutant approach. *Epidemiology* 2010;21:187–94.
2 Lenters V, Portengen L, Smit LA, *et al*. Phthalates, perfluoroalkyl acids, metals and organochlorines and reproductive function: a multipollutant assessment in Greenlandic, Polish and Ukrainian men. *Occup Environ Med* 2015;72:385–93.
3 Lenters V, Portengen L, Rignell-Hydbom A, *et al*. Prenatal phthalate, perfluoroalkyl acid, and organochlorine exposures and term birth weight in three birth cohorts: multi-pollutant models based on elastic net regression. *Environ Health Perspect* 2016;124:365–72.
4 Zanobetti A, Austin E, Coull BA, *et al*. Health effects of multi-pollutant profiles. *Environ Int* 2014;71:13–19.
5 Czarnota J, Gennings C, Colt JS, *et al*. Analysis of environmental chemical mixtures and Non-Hodgkin lymphoma risk in the NCI-SEER NHL Study. *Environ Health Perspect* 2015;123:965–70.
6 Chadeau-Hyam M, Campanella G, Jombart T, *et al*. Deciphering the complex: methodological overview of statistical models to derive OMICS-based biomarkers. *Environ Mol Mutagen* 2013;54:542–57.
7 Agier L, Portengen L, Chadeau-Hyam M, *et al*. A Systematic comparison of linear regression-based statistical methods to assess exposome-health associations. *Environ Health Perspect* 2016;124:1848–56.
8 Sun Z, Tao Y, Li S, *et al*. Statistical strategies for constructing health risk models with multiple pollutants and their interactions: possible choices and comparisons. *Environ Health* 2013;12:85.
9 Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Series B Stat Methodol* 1995;57:289–300.
10 Patel CJ, Bhattacharya J, Butte AJ. An Environment-Wide Association Study (EWAS) on type 2 diabetes mellitus. *PLoS One* 2010;5:e10746.
11 Wold S, Ruhe A, Wold H, *et al*. The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses. *SIAM J Sci Comput* 1984;5:735–43.
12 Chun H, Keleş S. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *J R Stat Soc Series B Stat Methodol* 2010;72:3–25.
13 Chung D, Keles S. Sparse partial least squares classification for high dimensional data. *Stat Appl Genet Mol Biol* 2010;9:Article17.
14 Tibshirani R. Regression shrinkage and selection via the Lasso. *J R Stat Soc Series B Stat Methodol* 1996;58:267–88.
15 Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc Series B Stat Methodol* 2005;67:301–20.
16 Hastie TJ, Tibshirani RJ, Friedman JH. *The elements of statistical learning: data mining, inference, and prediction*. 2nd ed. New York, NY: Springer, 2013.
17 Gramacy RB, Polson NG. Simulation-based Regularized Logistic Regression. *Bayesian Anal* 2012;7:567–90.
18 Park T, Casella G. The Bayesian Lasso. *J Am Stat Assoc* 2008;103:681–6.
19 Buhlmann P, Hothorn T. Regularization Boosting algorithms: prediction and model fitting. *Statistical Science* 2007;22:477–505.
20 Friedman J, Hastie T, Tibshirani R. Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors). *The Annals of Statistics* 2000;28:337–407.
21 Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One* 2015;10:e0118432.
22 Rijsbergen CJV. *Information retrieval*. 2nd ed. London: Butterworths, 1979.
23 Tibshirani RJ, Taylor J, Lockhart R, *et al*. Exact Post-Selection Inference for Sequential Regression Procedures. *J Am Stat Assoc* 2016;111:600–20.
24 G'Sell MG, Wager S, Chouldechova A, *et al*. Sequential selection procedures and false discovery rate control. *J R Stat Soc Series B* 2016;78:423–44.
25 Meinshausen N, Bühlmann P. Stability selection. *J Roy Stat Soc B* 2010;72:417–73.
26 Kyung M, Gill J, Ghosh M, *et al*. Penalized regression, standard errors, and Bayesian lassos. *Bayesian Anal* 2010;5:369–411.
27 Scott JG, Berger JO. Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *Ann Statist* 2010;38:2587–619.
28 Polson NG, Scott JG, Windle J. Bayesian inference for logistic models using Polya-Gamma latent variables. *J Am Stat Assoc* 2013;108:1339–49.
29 Bhadra A, Datta J, Polson NG, *et al*. The horseshoe+estimator of ultra-sparse signals. *Bayesian Anal* 2016 (EPub ahead of print: 22 Sep 2016).
30 Rockova V, Lesaffre E, Luime J, *et al*. Hierarchical Bayesian formulations for selecting variables in regression models. *Stat Med* 2012;31:1221–37.
31 Xing G, Lin CY, Xing C. A comparison of approaches to control for confounding factors by regression models. *Hum Hered* 2011;72:194–205.
32 Krstajic D, Buturovic LJ, Leahy DE, *et al*. Cross-validation pitfalls when selecting and assessing regression and classification models. *J Cheminform* 2014;6:10.
33 Bobb JF, Valeri L, Claus Henn B, *et al*. Bayesian kernel machine regression for estimating the health effects of multi-pollutant mixtures. *Biostatistics* 2015;16:493–508.
34 Austin E, Pan W, Shen X. Penalized regression and risk prediction in Genome-Wide Association Studies. *Stat Anal Data Min* 2013;6:315–28.
35 Patel CJ, Manrai AK, Development of exposome correlation globes to map out environment-wide associations. *Pacific symposium on biocomputing pacific symposium on biocomputing*, 2015:231–42.
36 Sinisi SE, van der Laan MJ. Deletion/substitution/addition algorithm in learning with applications in genomics. *Stat Appl Genet Mol Biol* 2004;3:1–38.
37 Bottolo L, Richardson S. Evolutionary stochastic search for Bayesian model exploration. *Bayesian Anal* 2010;5:583–618.
38 Dionisio KL, Baxter LK, Chang HH. An empirical assessment of exposure measurement error and effect attenuation in bipollutant epidemiologic models. *Environ Health Perspect* 2014;122:1216–24.
39 Pollack AZ, Perkins NJ, Mumford SL, *et al*. Correlated biomarker measurement error: an important threat to inference in environmental epidemiology. *Am J Epidemiol* 2013;177:84–92.
40 Groenwold RH, Sterne JA, Lawlor DA, *et al*. Sensitivity analysis for the effects of multiple unmeasured confounders. *Ann Epidemiol* 2016;26:605–11.
41 Schisterman EF, Perkins NJ, Mumford SL, *et al*. Collinearity and causal diagrams: a lesson on the importance of model specification. *Epidemiology* 2017;28:47–53.
42 Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw* 2010;33:1–22.
43 Hothorn T, Buhlmann P, Kneib T, *et al*. Model-based boosting 2.0. *J Mach Learn Res* 2010;11:2109–13.