How do you decide which variables to include? A real-world EMR case study comparing variable selection methods and their importance

Eleanor Ralphs[1] (eralphs@uk.imshealth.com), Peter McMahon[1], Benjamin Bray[1], Fiona Grimson[1] & Joseph Kim[1,2]

[1]NEMEA Centre of Excellence for Retrospective Studies, Real-World Insights, IQVIA, London, UK
[2]Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine

IMS Health & Quintiles are now IQVIA™

Aims

- To compare eight widely-used variable selection methods, by discussing the advantages and disadvantages, as well as the appropriate and inappropriate applications of the methods.
- To assess how the association between pregnancy and type II diabetes mellitus diagnosis differs when applying the eight different variable selections, using real-world EMR data.

Background

- Variable selection is the process of deciding which variables should be included in a statistical model. The use of an inappropriate selection of variables can lead to issues with **overfitting or confounding**, which can **misinform results.**
- The difficulty of knowing which variable selection method to use provides motivation for a **case study;** this will test the most common selection methods on the association between pregnancy and type II diabetes, with **real-world EMR** (electronic medical records). By understanding which variables need to be adjusted; statisticians will be able to more confidentially run regression models to provide more accurate results.

Description of variable selection methods

Online searches that yielded the most appropriate scientific articles were included in the tables below.

Method	Description
Bivariable Analysis[1,2]	• Bivariable analysis is conducted on a dependent and independent variable • If **the p-value** obtained from a test of association is significant, then it is included in the model • The method is often used to determine if a factor will be **considered for inclusion in the multivariate analysis**
F-test stepwise regression[3]	• An automated procedure, with a **bidirectional approach** to add and delete of variables in a model, where significance is defined by an arbitrary cut-off point for the **p-value** • Variables already in the model can be removed if they **become non-significant** after new variables are added
AIC (Akaike information criterion) stepwise regression[4]	• Similar to F-test stepwise regression (**bidirectional approach**) • Uses **AIC values**, which are the natural log of the likelihood function maximum of the model subtracted from the number of parameters • Assesses the difference between data generated by the model and the original data (estimates how much **information is lost**), those with the **lowest AIC** are retained for the next iteration
LASSO (least absolute shrinkage and selection operator)[5,6]	• LASSO regression is used for variable selection and L1 regularisation. L1 regularisation is when **large regression coefficients** which contribute to overfitting, are **set to zero** • LASSO decides which variables to shrink by putting a **constraint on the sum of the absolute values** of the model parameters • Variables which contribute to a **large sum of squared errors**, are not selected for the model
LARS (least-angle regression)[7]	• LARS makes **optimally-sized leaps** in the optimal direction 1. All coefficients are equal to **zero** 2. Finds the predictor **most correlated** with the response 3. Performs simple **linear regression** in the direction of the predictor, until another predictor has as much correlation with the current residual 4. LARS proceeds in an **equiangular direction** between the two predictors, until a third predictor becomes as correlated with the current residual 5. LARS then proceeds equiangularly between the three predictors, along the **'least angle direction'**, until a fourth predictor becomes highly correlated with the residual
All-subset selection[3]	• All **possible subsets** are created; a **univariate** model is generated for each variable, then all **bivariable** models, **three-variable** models etc • The R-squared, adjusted R-squared, BIC (Bayesian information criterion) or Mallows' Cp value are used as the criteria for variable selection
Forward selection[8]	• Variables are **continually added** to the crude model manually, only kept in if there is a considerable (10%) change in estimate • Would be less prone to inclusion if a variable considerably **increases the standard error** • Alternatively, the candidate variables with the largest change in **mean-square error** can be put in the model
Backward selection[1,8]	• The initial model has **all candidate variables** and variables which do not have a considerable (10%) **change in estimate** are manually removed from the model • The model is successively **refitted to be reduced**

Methods	Advantages	Disadvantages	Appropriate applications	Inappropriate applications
Bivariable analysis[1,2]	**Easy** to conduct on standard software	• **Important variables could be disregarded** if they appears to not be significantly associated with the outcome, as the association is **confounded** by another factor • Does not provide any benefits when building **multivariable models**	Recommended when there is likely to be **no confounding** factors	Not recommended when the variable and outcome is associated weakly, and **confounding** is likely
F-test stepwise regression[3]	Computationally **efficient**	• Multiple testing involved which can **overestimate the significance** of the remaining variables • **Bias regression coefficients** are produced which are falsely large • Very **narrow confidence intervals** are produced • Based on methods which test pre-determined hypotheses • **P-values** have debateable meaning • Models are prone to **overfitting** the data • **Collinearity** can be problematic • Depends on the **order of removal**	Recommended for **exploratory** data analysis	Not recommended to produce reliable results
AIC stepwise regression[4]	No reliance on p-value significance thresholds **Widely applied** and applicable for non-normally distributed data		Recommended for **exploratory** data analysis or s**impler models**	Not recommended to produce reliable results (although AIC is more favourable to other stepwise regression)
LASSO[5,6]	**Overfitting is reduced** as irrelevant variables have their regression coefficients set to zero Computationally **efficient** **Reduces variance** without substantially increasing bias	If regularisation is too strong, coefficients can be **excessively shrunk**, and important variables can be left out	Recommended when there are **many candidate variables** present (recommended to use the LASSO method over stepwise approaches)	Not recommended when it does **not make sense to zero** large regression coefficients in your dataset
LARS[7]	Fits linear regression models to high-dimensional data **Avoids overfitting**	LARS is specifically **sensitive** to the effect of noise	Recommended for **high-dimension** dataset	Not recommended when variables are **highly correlated**
All-subset selection[3]	Models with **all variable combinations** are produced	• Produces **many models**, needs sufficient computational power • For large models (>40 variables) the regression statistics and **coefficient are bias**	Recommended when **few candidate variables** present	Not recommended when a quick and efficient approach is desired
Forward selection[8]	The confounding effect of **each** additional variable can be seen	• Forward selection fails to identify if a set of variables **collectively influence** the significance of the model • **Time consuming**	Recommended when there are **more candidate variables** to consider than can reasonably be fitted at once	Not recommended when several non-significant **variables jointly behave**
Backwards elimination[1,8]	This approach allows for a set of variables to have considerable **predictive capability collectively**, even though individually they may not	• Depends on the **order of removal** • **Collinearity** is more likely to occur • The approach is problematic when variables have **sparse data**	Recommended when there is a **collectively significant** effect between variables and when using **prior knowledge** to select variables	Not recommended when several variables are **correlated**

Case study: Association between pregnancy and type II diabetes

Methods

- Data source: **IQVIA Medical Research Data**. This is a large UK primary-care database comprised of over 14 million patients, containing electronic medical record information including diagnoses, prescribed therapies, tests and referrals. The September 2018 version of the data was extracted using the E360 Cohort Builder platform.
- Study design: This is a **retrospective cohort study**. Looking at the association between pregnancy and type II diabetes mellitus.
- Study population: The study population was limited to females ≥18 years of age. A random sample of 10% of the available patients was extracted, yielding a study population size of **216,039 patients**.
- Study time period: 1st April 2014 – September 2018

Analysis

- Each of the eight variable selection methods were run on the extracted data. Once the covariables were selected (table 1), **multivariable binary logistic regressions** were performed on the association between pregnancy and type II diabetes. The results were compared in a forest plot (figure 1). All analysis was conducted in R version 3.5.1.

Table 1. Comparison of variables selected, using eight different variable selection methods

All variables extracted	Bivariable analysis	F-test stepwise	AIC stepwise	LASSO	LARS	All-subset selection	Forward selection	Backward selection
Age	X	X	X	X	X	X	X	X
BMI	X	X	X	X	X	X		X
Obesity	X	X	X	X	X	X		
Insulin resistance	X	X	X	X	X	X	X	
Hypertension	X	X	X	X	X	X		
Dyslipidaemia	X	X	X	X	X	X		
Cardiovascular disease	X	X	X	X				
Current smoker	X	X	X					
Ex-smoker	X	X	X					
Non-smoker		X	X					
Cystic Fibrosis			X					
Acromegaly			X					
Cushing's disease			X					
Cardiac arrest			X					
Alcohol abuse								

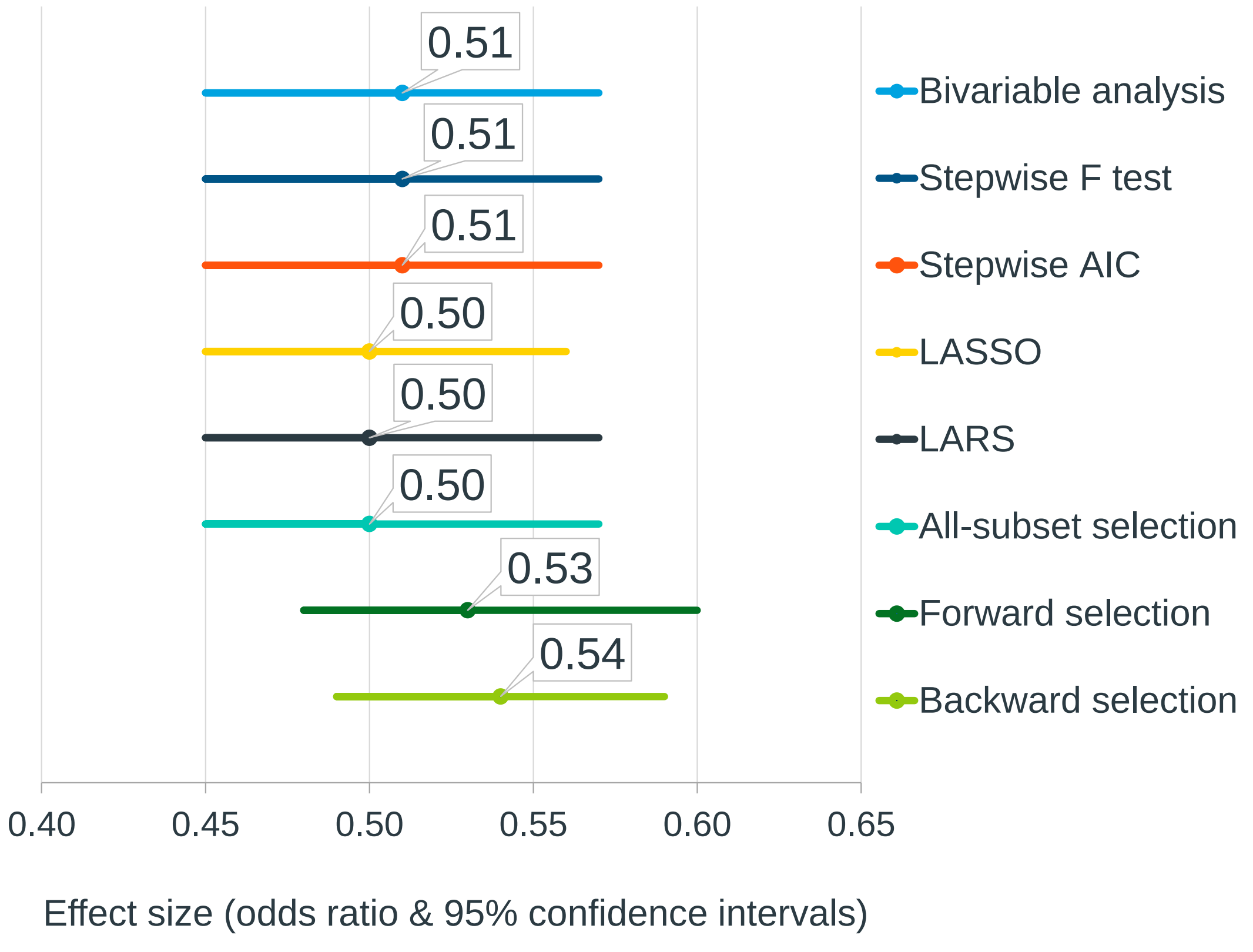
Colour key: X Dark grey= demographics variables, X grey= clinical variable, X light grey= behavioural variables

Choice of variables

It is evident that there is variation in variable choice, between the different selection methods.

- **Stepwise methods:** chose for the most variables. A disadvantage of these methods is their inclination to **overfit** the data. It appears that the stepwise methods could be overfitting due to including variables that the other methods deem redundant.
- **Forward and backward manual selection:** chose for the fewest variables, which could have been caused by **collective confounding** between variables not being picked up, due to the order of variable removal during the selection process.

Figure 1. Forest plot comparing odds ratios produced using eight different variable selection methods



Comparison of results yielded by the different selection methods

All outcomes from the eight different models suggest that **women who have not been pregnant have approximately half the odds of being diagnosed with type II diabetes,** compared to women who have been pregnant. However this data carries some caveats, explained in the limitations. The case study is intended to compare selection methods and not intended to make medical inferences.

Each selection method produced the **same odds ratio** within a value of 0.04, and all had p-values <0.001. The confidence intervals are approximately similar widths, and the standard errors are all very similar and small, being 0.05-0.06.

The forward and backward selection manual methods produced the odds ratios most varied from the mean of the other methods. This could be due to the considerably fewer variables adjusted for in their final models.

In this case, **LASSO would be the most appropriate** selection method, as the data contains many candidate variables and the advantages of LASSO outweighs those of other selection methods available.

Limitations

- A limitation of the online research includes the potential for a **publication bias** regarding the type of articles available. Critiques of more common methods may be more likely to be reported.
- The case study is limited by the data source and the non-clinical approach. IQVIA Medical Research Data is a representative and generalisable database, comprised of approximately 5% of the UK population. However, the representativeness of this study population is unknown, so the results have **limited generalisability.**
- Additionally, the true association between pregnancy and type II diabetes may not be presented here, due to potential **residual confounding.** There is a lack of prescription and lifestyle information (e.g. alcohol abuse, lack of exercise) available which could have provided a more precise result. Also, due to the focus being on comparing selection methods, rather than the medical inference, no consideration on timing of pregnancy and type II diabetes was taken into account; the women could have been diagnoses with type II diabetes before pregnancy. Additionally those flagged as pregnant could include patients who had terminated pregnancies.
- **Further work** would explore more case studies which tests the different variable selection methods, using real-world EMR data. This would determine if the same conclusions are reached, irrespective of the disease area and association of interest.

Conclusions

From the research

- With **different data types,** different selection methods can be most appropriate
- LASSO and LARS are highly **preferred** by the literature, presenting fewer disadvantages
- Stepwise regression methods are the most **criticised** in the literature

From the case study

- Irrespective of which selection methods is used, the **effect estimates and interpretations are very similar** in the case study presented
- Stepwise methods selected for most variables, whereas forward and backward manual methods selected for fewest variables
- In this case study, **LASSO would be the most appropriate** selection method, as the data contains many candidate variables and the advantages of LASSO outweighs those of other selection methods available

References

1. Heinze G, Dunkler D. Five myths about variable selection. Transplant International. 2017 Jan;30(1):6-10.
2. Sun GW, Shook TL, Kay GL. Inappropriate use of bivariable analysis to screen risk factors for use in multivariable analysis. Journal of clinical epidemiology. 1996 Aug 1;49(8):907-16.
3. Ratner B. Variable selection methods in regression: Ignorable problem, outing notable solution. Journal of Targeting, Measurement and Analysis for Marketing. 2010 Mar 1;18(1):65-75.
4. Yamashita T, Yamashita K, Kamimura R. A stepwise AIC method for variable selection in linear regression. Communications in Statistics—Theory and Methods. 2007 Oct 3;36(13):2395-403.
5. Sartori S. Penalized regression: Bootstrap confidence intervals and variable selection for high-dimensional data sets. PhD thesis, Universitá degli Studi di Milano, 2011.
6. Fonti V, Belitser E. Feature selection using lasso. VU Amsterdam Research Paper in Business Analytics. 2017 Mar 30.
7. Efron B, Hastie T, Johnstone I, Tibshirani R. Least angle regression. The Annals of statistics. 2004 Apr;32(2):407-99.
8. Guyon I, Elisseeff A. An introduction to variable and feature selection. Journal of machine learning research. 2003;3(Mar):1157-82.

IQVIA Medical Research Data is a registered trademark of Cegedim SA in the United Kingdom and other countries. Reference made to IQVIA Medical Research Data is intended to be descriptive of the data asset licensed by IQVIA. This work uses de-identified data provided by patients as a part of their routine primary care.