**Reducing and Quantifying Over-Fitting in Regression Models in Healthcare Studies**

# John Rigg1\*¶, Matthew Hankins1¶

# 

# 1 Real-World Evidence Solutions, IMS Health, London, UK

\*Corresponding author

# Email: [john.rigg@uk.imshealth.com](mailto:john.rigg@uk.imshealth.com)

# ¶ These authors contributed equally to this work.

# Abstract

...

# Introduction

The application of regression models to derive estimates of risk factors is one of the most common analytical tasks carried out by epidemiologists and statisticians working with healthcare data. It is the principal multivariable analytical approach employed to understand risk factors associated with disease prevalence, disease progression, medication adherence, healthcare utilisation and a myriad of other facets of medical empirical science. It is widely-recognised that these models may experience over-fitting, i.e. the results of the study fail to generalise beyond the data used to drive the estimates of risk. Over-fitting is potentially an extremely serious problem since it may completely invalidate results of a study.

Despite this, studies seldom explicitly attempt to minimise over-fitting beyond limiting the number of covariates to enter the model in the first place. Moreover, tests are almost never carried out to assess out-of-sample accuracy and hence the extent to which models may suffer from over-fitting. This reflects a reluctance to hold data back for independent post-modelling validation, fuelled by a desire to use all available data for model estimation. The problem of over-fitting is likely to become more acute as the number of potential covariates increase with the availability of more detailed patient-level data (e.g. from biomarker information and combined data from multiple care settings).

The aims of this study are to:

* Examine the extent to which standard healthcare studies may experience over-fitting.
* Illustrate the extent to which over-fitting may produce misleading conclusions, arising from inflated model parameters, false-positive inferences and overly optimistic evaluation of model performance.
* Assess the effectiveness of penalized regressions for reducing over-fitting in typical healthcare studies. Penalised (or regularised) regressions impose a constraint on the optimisation task resulting in smaller, potentially less inflated coefficients.
* Evaluate the accuracy of cross-validation for estimating over-fitting without using hold-out data. Cross-validation is a resampling method which involves splitting data into separate samples for model estimation and assessment.

Resampling methods (such as cross-validation) to quantify over-fitting and penalised methods to reduce over-fitting are both well-established methods in statistical learning and machine learning.1-3 The principal aim of this study is to apply these methods in a typical healthcare setting where they are rarely employed; the results can inform on the extent to which typical healthcare studies may experience over-fitting and inflated risk factors and hence whether these methods can help address such problems. A ‘typical’ healthcare study was carefully designed, with analysis based on multiple regression methods, outcomes, covariates and data, all of which are routinely employed in healthcare studies. Whilst informative, other scenarios and datasets are required to verify the extent to which these findings generalise.

Penalised regression involves minimising the sum of squared residuals plus a penalty term. The magnitude of the penalty is a parameter to be optimised. In effect, the regularisation penalises model complexity. The penalty can take several forms (discussed below). LASSO penalised regression4 is implemented in this study. LASSO shrinks some of the coefficients to zero, hence it is also a method of variable selection. LASSO is chosen as the form of penalisation in this study since variable selection is a common objective of many studies (as evidenced, for example, by the widespread use of stepwise regressions). Recent examples of LASSO regressions include identification of risk factors for mortality after myocardial infarction5 development and validation of a stroke risk model6 prediction of thirty-day hospital readmissions.7 Prediction of health-related quality of life in COPD patients8 and prognostic risk scoring for survival for metastatic castration-resistant prostate cancer.9

Compared to standard regression methods, a distinguishing feature of regularised regression is the use of cross-validation to optimise model performance. Cross-validation involves randomly splitting cases into K exhaustive and mutually exclusive folds (segments). A common value for K is ten since this has been shown to have good bias and variance properties.2,10 Assuming ten folds are selected, a model with a given penalty value (constraint) is estimated on nine of the ten folds and predictions computed for the tenth, left-out fold. Ten models are estimated with a different fold left-out for each run. Model accuracy is then calculated using an appropriate metric such as the residual sum of squares. This provides an estimate of out-of-sample accuracy since the predictions used to calculate the performance metric are all based on out-of-sample data. Different penalty values are applied to the regression and cross-validation used to assess model performance. The model with the constraint associated with the best performance metric is selected. The use of an out-of-sample performance metric to assess model fit is a key aspect of penalised regression which helps explain any improvement in out-of-sample accuracy and over-fitting compared to conventional implementations of generalised linear models.

In addition, cross-validation is a resampling method which can be used to provide an estimate of out-of-sample model accuracy. This is obviously of considerable value since it informs on the extent to which the results may be generalised. It also enables the extent of over-fitting to be estimated without holding-back any of the data. Armed with this estimate of out-of-sample model accuracy, it is then possible to take steps design to improve the generalisation and over-fitting properties of the model.

This study assesses the use of penalised regressions and cross-validation for reducing and quantifying over-fitting. Results are compared between standard (unconstrained), stepwise and LASSO logistic regressions. The standard and stepwise logistic regression methods are chosen since they are two of the most widely adopted regression approaches where outcomes are dichotomous. The impact of over-fitting is illustrated by comparing the magnitude of odds ratios for key regressors for different models varying by the degree of over-fitting.

The methods are assessed using retrospective data on 3,348 patients with Relapse Remitting Multiple Sclerosis. All patients receive one of two disease-modifying therapies. Data are extracted from IMS Pharmetrics Plus retrospective adjudicated medical claims database.

The sensitivity of findings is assessed with respect to the choice of response variable and the degrees of freedom. In particular, separate models are estimated for two binary outcomes: whether the patient experienced a relapse and whether the patient was persistent on medication. There were 46 covariates entered into each model. Baseline estimates are provided using 1,674 cases to train the models and the impact of reducing the degrees of freedom illustrated by re-estimating the models using just 400 patients. Constraining the degrees of freedom is expected to lead to an increase in over-fitting. The ratio of 8.7 cases per candidate covariate (400/46) is low but far from unprecedented. It is therefore instructive to assess over-fitting in this setting and the extent to which regularisation can help mitigate the problem.

# Patients and Methods

## Data

Data is used for two common treatments for Relapse Remitting Multiple Sclerosis, referred to as treatment A and treatment B. The data is extracted from IMS Pharmetrics Plus retrospective database of adjudicated medical claims. This covers a large representative sample of the US commercially-insured population

The index date is the first observed date receiving treatment A or B. All patients had at least twelve months pre-index and six months post-index health plan enrollment.

Separate analysis is carried out for two binary response variables. The first indicates whether the patient relapsed in the first six months post-index; the second indicates whether the patient was persistent to medication for the first six months post-index.

Covariates captured the following dimensions:

* Treatment: Whether the patient received treatment A or B at index date.
* Healthcare utilisation: This includes pre-index medication use and the number of pre-index MS-related visits and procedures.
* Healthcare costs: This separately covers total pre-index MS-related and non-MS-related costs.
* Relapse history: Whether the patient experienced a relapse in the twelve months prior to index.
* Pre-index co-morbidities: This covers a large set of co-morbidities, many of which are common amongst MS patients.
* Demographics: Gender, region and age.
* Index prescribing physician specialty e.g. neurology, general practice, etc.
* Index provider type e.g. Health Maintenance Organisation, consumer directed, etc.

Patients receiving treatments A and B were matched using propensity scoring11 to balance differences in pre-index confounders. There are 1,674 matched pairs (i.e. 3,348 patients in total) receiving either treatment A or B. Patients were matched by strata, defined as 0, 1 or 2+ relapses in the six months post-index period. Patients receiving treatment A and B were matched one-to-one within strata.

In keeping with standard practice, (unconstrained) logistic regressions were used to compute the propensity scores. A covariate was considered to be balanced if the absolute standardised difference in means between treatmen groups was less than 10% (a commonly applied threshold.12 The following covariates were balanced: Age, gender, region, physician specialty, pre-index use of MS disease modifying therapy, pre-index use of ampyra, pre-index MRI use, timing of pre-index costs (MS and non-MS) and key co-morbidities.

## Methods

This study applied Standard (unconstrained), (backward) stepwise and LASSO logistic regressions. The sensitivity of findings was assessed with respect to the choice of response variable and the degrees of freedom. In particular, standard (unconstrained), backwards stepwise and LASSO logistic regressions were estimated for two response variables: whether the patient experienced a relapse post-index and whether the patient was persistent to medication in the post-index period. For each model, the data was randomly split 50/50 into training and test data (i.e. the training sample comprised 1,674 cases and the test sample comprised 1,674 cases). To minimise noise arising from random patient sample asignment, 100 iterations of each model type were estimated and the mean results were reported. To assess the sensitivity of results to the degrees of freedom, all models were rerun where just 400 patients were randomly assigned to the training set. This is a ratio of cases to covariates of 8.7 (400/46) which is unusual but far from unprecedented.

Given the understandable and widespread reluctance to hold back data and thereby reduce the available data for training purposes, an important question is whether training data alone can provide an accurate assessment of out-of-sample model accuracy and hence the extent of over-fitting. Estimates of model performance on test data were therefore computed using tenfold cross-validation on training data only. The estimates of model performance on test data were then compared to actual model performance on test data.

The primary measure used to assess model performance was the Area Under Curve (AUC) for the graph of false positive vs. true positive rate for each model prediction of outcome.13 Over-fitting was calculated as the difference between the AUC on training and test data.

A review of penalised regressions is available in Hesterberg et al. (2008)14 and Hastie et al. (2009).2 Non-technical introductions including worked examples (with R code) are available in James et al (2013)15 and Kuhn and Johnson (2013).16

Consider a standard multivariate regression model where *y*t is estimated as a linear function of a constant, *b*0, and *n* covariates. Assume all (non-constant) covariates are normalised to have mean zero and variance one. Consider selecting the coefficients *b*1,…,*b*n for these covariates by minimizing the sum of squared residuals plus a penalty term of the form:17

This estimation method is known as elastic net regression.18 It contains three methods as special cases: if there is no penalty term (λ = 0) this is ordinary least squares regression; if alpha = 1, so that there is only the quadratic constraint, this is ridge regression.19 If alpha = 0, this is known as the LASSO, an acronym for least absolute shrinkage and selection operator.4

Over-fitting is characterised by inflated model parameters. Penalised regression constrains the magnitude of coefficients, shrinking the least squares regression coefficients towards zero (these methods are therefore also known as ‘shrinkage’ and ‘sparse’ regression). The LASSO and elastic net typically produce regressions where some of the variables are set to be exactly zero. Hence this is also a means of variable selection. While ridge regression shrinks the parameter estimates towards 0, the model does not set the values to absolute 0 for any value of the penalty. Even though some parameter estimates become negligibly small, this model does not therefore conduct variable selection.

Computing confidence intervals for coefficients obtained from LASSO regressions is an active area of research.20 The challenge arises since the distribution of parameters for the null hypothesis is unknown (the conventional normality assumption is violated). In this study, the odds ratios kept by the LASSO regressions were reported along with p-values obtained from a standard logistic regression including only the variables with non-zero coefficients retained by the LASSO regression (for a similar approach, see Halabi et al., 2013 9).

The LASSO approach used in this study has been extended to many techniques, such as linear discriminant analysis,21-22 Partial Least Squares23 and Principal Components Analysis.24-25

One common setting where penalised regressions perform well is where there is multicollinarity amongst covariates. Multicollinearity often produces models with high variance; models with modest bias can lead to substantially lower variance and hence lower overall error.26 Importantly, penalised regressions can be computed efficiently27 so carrying out variable selection on large datasets (e.g. genomics data) is computationally feasible.

# Results

## Descriptive Statistics

Descriptive statistics are reported in Table 1. Frequency counts for the response variables showed that 428 patients (12.8%) experienced a relapse and 2,669 patients (79.7%) were persistent in the six months post-index period. The basic demographic variables showed that the sample was predominantly female (2,585 women to 763 men) and the average patient was 45.6 years old. The sample comprised a mix of payer types and provider organisations and, as expected, relatively high counts of healthcare use and co-morbidities (e.g. 546 patients reported muscle weakness, spasm or spasticity in the 12 months pre-index period). 1,254 patients experienced at least one relapse in the pre-index period.

Table 1. Descriptive statistics by treatment.

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** |  | **Treatment**  **A** | **Treatment**  **B** |
| Sex (Female) | N | 1287 | 1298 |
| Index Age | Mean (SD) | 45.7 (10.6) | 45.5 (9.9) |
|  | Min-Max | 18-73 | 18-72 |
| Pre-index DMT use | N | 1180 | 1187 |
| **Region** | | | |
| Northeast | N | 483 | 479 |
| Midwest | N | 556 | 584 |
| South | N | 506 | 485 |
| **Index Payer type** | | | |
| Commercial | N | 1129 | 1106 |
| Self-insured | N | 520 | 545 |
| **Index Provider type** | | | |
| Health Maintenance Organization | N | 157 | 221 |
| Preferred Provided Organization | N | 1363 | 1324 |
| Point of Service | N | 83 | 79 |
| **Index Prescribing physician specialty** | | | |
| General practice/Family practice | N | 104 | 84 |
| Internal medicine | N | 63 | 75 |
| Neurology | N | 832 | 837 |
| Other | N | 635 | 630 |
| **Pre-index symptoms /comorbidities** | | | |
| Numbness | N | 373 | 368 |
| Walking (Gait), Balance, and Coordination Problems | N | 365 | 355 |
| Dizziness and Vertigo | N | 182 | 174 |
| Muscle weakness/spasm/spasticity | N | 282 | 264 |
| Fatigue | N | 527 | 524 |
| Bladder Dysfunction | N | 230 | 207 |
| Bowel Dysfunction | N | 188 | 160 |
| Visual Symptoms | N | 243 | 237 |
| Pain | N | 114 | 98 |
| Seizures | N | 48 | 66 |
| Tremor | N | 90 | 93 |
| Respiration/breathing problems | N | 161 | 159 |
| Depression Comorbidity | N | 371 | 351 |
| Diabetes mellitus | N | 117 | 111 |
| Dyslipidemia | N | 405 | 415 |
| History of CVD | N | 88 | 106 |
| Obesity | N | 102 | 81 |
| Tobacco use (including disorder) | N | 148 | 142 |
| **Pre-index** | | | |
| Ampyra use | N | 166 | 149 |
| MRI use | N | 1263 | 1270 |
| Corticosteroid use within 90 days pre-index | N | 442 | 423 |
| Oral corticosteroid use | N | 590 | 531 |
| Iv corticosteroid use | N | 569 | 573 |
| Presence of pre-index relapse | N | 627 | 627 |
| **Variable** |  | **Treatment**  **A** | **Treatment**  **B** |
| No. of unique pre-index medications | N | 872 | 843 |
| No. of pre-index OP visits for MS diagnosis | N | 826 | 1012 |
| Total non-MS pre-index costs | N | 841 | 833 |
| MS-related total pre-index costs | N | 902 | 772 |
| Charlson Co-morbidity Index = 1 | N | 263 | 264 |
| Charlson Co-morbidity Index = 2+ | N | 235 | 231 |

SD, standard deviation; DMT, **Dimethyltryptamine**; CVD, Cardiovascular disease; MRI,magnetic resonance imaging ; OP, outpatient; MS, multiple sclerosis.

## Relapse

Table 2 reports results for both in-sample and out-of-sample model accuracy and the amount of over-fitting for each model with relapse as the predicted variable. For the model using 1,674 cases each for training and test data, the AUC on training data was 74.4 for the standard logistic regression and 68.0 for the test data, representing over-fitting of 9.9%. The stepwise procedure was successful in achieving a modest reduction in over-fitting to 8.0%. The LASSO further reduced over-fitting to 3.9%. The improvement associated with the LASSO was due to both a lower AUC on training data and increased classification accuracy on test data. This latter point is important since it highlights that LASSO produced the most accurate results on out-of-sample data which in many contexts is the single most important consideration. The AUC on test data was 68.0 for both the standard and stepwise logistic regressions and 69.5 for the LASSO.

Table 2. Accuracy and overfitting by model type (relapse).

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Model | N for training sample | N for test sample | AUC for training sample | SD for AUC on training sample | AUC for test sample | SD for AUC on test sample | Overfitting(absolute difference) | Overfitting(% difference) |
| Standard Logistic | 1,674 | 1,674 | 74.7% | 1.3% | 68.0% | 1.5% | 6.7% | 9.9 |
| Stepwise Logistic | 1,674 | 1,674 | 73.5% | 1.4% | 68.0% | 1.4% | 5.5% | 8.0 |
| Lasso | 1,674 | 1,674 | 72.2% | 1.8% | 69.5% | 1.5% | 2.7% | 3.9 |
| Model | N for training sample | N for test sample | AUC for training sample | SD for AUC on training sample | AUC for test sample | SD for AUC on test sample | Overfitting(absolute difference) | Overfitting(% difference) |
| Standard Logistic | 400 | 2,948 | 82.7% | 2.8% | 62.3% | 2.5% | 20.4% | 32.7 |
| Stepwise Logistic | 400 | 2,948 | 79.4% | 3.3% | 63.4% | 2.5% | 16.0% | 25.1 |
| Lasso | 400 | 2,948 | 76.5% | 4.6% | 66.6% | 2.1% | 9.9% | 14.8 |

AUC, area under curve; SD, standard deviation.

As expected, reducing the degrees of freedom by using just 400 cases in the training sample led to a marked increase in the extent of over-fitting as well as lower out-of- sample accuracy: over-fitting increased to 32.7% for the standard logistic regression ,

25.1% for the stepwise procedure and 14.8% for the LASSO procedure. Although there remained considerable over-fitting with LASSO, it was notably lower than with the alternative methods.

For a subset of variables for the models estimated on 1,674 observations, Table 3 shows the number of times each variable appeared as a significant predictor in each of the models, with the mean and standard deviations of the estimates obtained (the complete set of results for all models are reported in supporting information tables). Treatment was a significant predictor of relapse in 11% of logistic regression models with a mean odds ratio of 0.81. In other words, Treatment A was associated with a reduced chance of relapse compared to Treatment B in 11 of the 100 standard logistic models. The treatment covariate was retained in 46% of stepwise models and was significant in 16 final models with a mean odds ratio of 0.74. Finally, treatment was retained in 43% of lasso models, remaining significant in 10 models with a mean odds ratio of 0.90.

The effect of treatment reported in Table 5 was less equivocal than for relapse, with all models retaining treatment 100% of the time. However, while logistic and stepwise models estimated the mean odds ratio >2.0 for treatment effect, the mean odds ratio for lasso regression was a more modest 1.64, a reduction in the estimated treatment effect of around 20%, although clearly this may also be an over-estimate given that the LASSO was found to over-fit by around 4% (Table 3).

## AUC Computed Using Training Data Alone

Tables 2 and 3 report AUCs computed on test data, i.e. data which were held back from the estimation sample. Hence, these AUCs show the performance of the final model when used on new (unseen) data. Table 6 compares these results with estimates of test (out-of-sample) AUC, computed using tenfold cross-validation on training data only.

The estimated test AUCs provided a reasonable approximation of actual test AUCs. For instance, the actual AUC on test data was 68.0 for the standard logistic regression where the outcome was relapse and 1,674 observations were used each for training and testing. The estimate of test AUC for this model was 67.7 based only on the 1,674 training cases using standard logistic regression and tenfold cross-validation. In three out of four situations, the difference between the estimated and actual AUCs was less than or equal to 0.5. Even in the least accurate setting (modelling relapse using 400 observations for training), the estimated AUC provided a very good gauge of out-of-sample accuracy: the estimated test AUC was 60.4 compared to the actual test AUC of 62.3. This is a far better guide to out-of-sample accuracy than the AUC of 82.7 based on training data which is the result that studies would typically report.

Finally, given that the results in Table 6 show that cross-validation can provide a reasonable approximation of out-of-sample model accuracy and over-fitting, it is informative to assess model performance using all available observations (i.e. without holding back any data for testing). Thus, the standard logistic regressions were re-estimated on the full sample (N=3,348) and tenfold cross-validation used to estimate model accuracy and over-fitting. The larger sample enables a higher estimate of out-of- sample accuracy for both relapse and persistence and the estimated amount of over-fitting falls to 5.0% and 6.6% for relapse and persistence respectively (actual overfitting was 9.9% and 12.0% respectively when the models were estimated on half the sample, N=1,674). LASSO regressions were also estimated on the full sample (N=3,348) which

**Table 3. Odds ratios for 50/50 training/test samples (relapse).**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Variable Description** | **Standard logistic** | | | **Stepwise** | | | | **Lasso** | | | | **Model comparison** | | | |
|  | **No. of times significant** | **Mean OR** | **SD OR** | **Number of times retained** | **Number of times significant** | **Mean OR** | **SD OR** | **Number of times retained** | **Number of times significant** | **Mean OR (A)** | **SD OR** | **Mean OR in standard LR when retained by Lasso LR (B)** | **Difference in mean OR (A-B)** | **Mean OR in stepwise LR when retained by stepwise & Lasso (C)** | **Difference in mean OR (A-C)** |
| **Treatment** | 11 | 0.81 | 0.08 | 46 | 16 | 0.74 | 0.05 | 43 | 10 | 0.90 | 0.07 | 0.77 | 0.14 | 0.74 | 0.17 |
| **Index age** | 4 | 0.99 | 0.01 | 25 | 7 | 0.99 | 0.00 | 93 | 9 | 0.99 | 0.00 | 0.99 | 0.00 | 0.99 | 0.01 |
| **Corticosteroid use within 90 days pre-index** | 68 | 1.55 | 0.24 | 88 | 75 | 1.67 | 0.20 | 95 | 70 | 1.39 | 0.19 | 1.58 | -0.19 | 1.67 | -0.28 |
| **Presence of Pre-index relapse** | 7 | 1.30 | 0.21 | 48 | 31 | 1.65 | 0.21 | 94 | 18 | 1.36 | 0.19 | 1.32 | 0.03 | 1.65 | -0.29 |
| **Pre-index dyslipidemia** | 0 | 1.08 | 0.14 | 6 | 0 | 1.14 | 0.33 | 7 | 0 | 1.02 | 0.09 | 1.18 | -0.16 | 1.16 | -0.13 |
| **History of CVD** | 28 | 0.60 | 0.15 | 45 | 14 | 0.52 | 0.08 | 6 | 3 | 0.84 | 0.05 | 0.43 | 0.41 | 0.47 | 0.37 |
| **Charlson Comorbidity Index = 2+** | 16 | 1.37 | 0.27 | 30 | 19 | 1.67 | 0.23 | 14 | 4 | 1.17 | 0.15 | 1.77 | -0.60 | 1.84 | -0.67 |

OR, odds ratio: SD, standard deviation; LR, logistic regression; CVD, cardiovascular disease.

## Table 4. Accuracy and overfitting by model type (persistence).

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Model | N for training sample | N for test sample | AUC for training sample | SD for AUC on training sample | AUC for test sample | SD for AUC on test sample | Overfitting(absolute difference) | Overfitting(% difference) |
| Standard Logistic | 1,674 | 1,674 | 67.1% | 1.0% | 59.9% | 1.1% | 7.2% | 12.0 |
| Stepwise Logistic | 1,674 | 1,674 | 65.6% | 1.2% | 60.2% | 1.2% | 5.4% | 8.9 |
| Lasso | 1,674 | 1,674 | 63.9% | 1.4% | 61.3% | 1.0% | 2.6% | 4.2 |
| Standard Logistic | 400 | 2,948 | 75.1% | 2.3% | 55.9% | 1.6% | 19.2% | 34.3 |
| Stepwise Logistic | 400 | 2,948 | 71.3% | 3.2% | 56.8% | 2.4% | 14.5% | 25.4 |
| Lasso | 400 | 2,948 | 68.2% | 6.2% | 57.2% | 3.2% | 11.0% | 19.2 |

## AUC, area under curve; SD, standard deviation.

**Table 5. Odds ratios for 50/50 training/test samples (persistence).**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Variable Description** | **Standard logistic** | | | **Stepwise** | | | | **Lasso** | | | | **Model comparison** | | | |
|  | **Number of times significant** | **Mean OR** | **SD OR** | **Number of times retained** | **Number of times significant** | **Mean OR** | **SD OR** | **Number of times retained** | **Number of times significant** | **Mean OR (A)** | **SD OR** | **Mean OR in standard LR when retained by Lasso LR (B)** | **Difference in mean OR (A-B)** | **Mean OR in stepwise LR when retained by stepwise & Lasso (C)** | **Difference in mean OR (A-C)** |
| **Treatment** | 100 | 2.04 | 0.18 | 100 | 100 | 2.03 | 0.17 | 100 | 100 | 1.64 | 0.17 | 2.04 | -0.41 | 2.03 | -0.39 |
| **Index age** | 4 | 1 | 0.01 | 13 | 4 | 0.99 | 0 | 95 | 1 | 1 | 0 | 1 | 0 | 0.99 | 0.01 |
| **Corticosteroid use within 90 days pre-index** | 0 | 1.05 | 0.15 | 11 | 0 | 1.13 | 0.29 | 8 | 0 | 1.01 | 0.1 | 1.01 | 0 | 1.06 | -0.05 |
| **Presence of Pre-index relapse** | 24 | 1.37 | 0.21 | 47 | 25 | 1.45 | 0.16 | 11 | 7 | 1.14 | 0.08 | 1.65 | -0.51 | 1.52 | -0.38 |
| **Pre-index dyslipidemia** | 6 | 1.16 | 0.13 | 20 | 3 | 1.30 | 0.05 | 12 | 4 | 1.08 | 0.08 | 1.27 | -0.19 | 1.33 | -0.24 |
| **History of CVD** | 0 | 1.06 | 0.23 | 4 | 0 | 1.15 | 0.56 | 1 | 0 | 1.01 | NA | 1.37 | -0.35 | 1.52 | -0.50 |
| **Charlson Comorbidity Index = 2+** | 6 | 0.86 | 0.15 | 23 | 9 | 0.71 | 0.05 | 18 | 5 | 0.91 | 0.07 | 0.71 | 0.20 | 0.69 | 0.22 |

OR, odds ratio: SD, standard deviation; LR, logistic regression, CVD cardiovascular disease.

# Table 6. Using cross-validation to estimate out-of-sample accuracy and overfitting.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Outcome** | **Model type** | **N for training sample** | **N for test sample** | **AUC for training sample** | **AUC for test sample** | **Overfitting (% difference): AUC on training -AUC on test as % of test AUC** | **Estimate of AUC for test sample** | **SD of estimate of AUC for test sample** | **Accuracy of estimate: AUC for test -estimate of AUC for test sample** | **Estimated overfitting (% difference): AUC for training -estimate of AUC for test as % of estimate of AUC for test** |
| Relapse | Standard logistic | 1,674 | 1,674 | 74.7% | 68.0% | 9.9 | 67.7% | 1.8% | 0.3% | 10.4 |
| Relapse | Standard logistic | 400 | 2,948 | 82.7% | 62.3% | 32.7 | 60.4% | 5.7% | 2.0% | 37.0 |
| Relapse | Standard logistic | 3348 | NA | 73.0% | NA | NA | 69.3% | 0.3% | NA | 5.4 |
| Relapse | Lasso | 3348 | NA | 71.5% | NA | NA | 70.3% | 0.2% | NA | 1.7 |
| Persistence | Standard logistic | 1,674 | 1,674 | 67.1% | 59.9% | 12.0 | 59.6% | 1.5% | 0.3% | 12.6 |
| Persistence | Standard logistic | 400 | 2,948 | 75.1% | 55.9% | 34.3 | 56.4% | 4.9% | -0.5% | 33.1 |
| Persistence | Standard logistic | 3348 | NA | 65.3% | NA | NA | 61.0% | 0.3% | NA | 7.0 |
| Persistence | Lasso | 3348 | NA | 63.7% | NA | NA | 61.8% | 0.2% | NA | 3.1 |

AUC, area under curve; SD, standard deviation.

led to improvements in out-of-sample accuracy and over-fitting. For instance, the amount of over-fitting fell to 1.7% for relapse and 3.7% for persistence.

# Discussion

It is widely recognized that regression models may suffer from over-fitting, potentially compromising the generalisability of study findings. Despite this, there is little systematic attempt to minimize and quantify over-fitting.

This study was designed to replicate the methods and data commonly encountered in many healthcare studies. However, there is considerable diversity in research materials and methods across studies; the conclusions in this study would benefit from replication in other settings to establish the extent to which these findings generalise.

This study found that standard and stepwise logistic regressions were associated with considerable over-fitting of approximately 9-35%, with the extent of over-fitting inversely associated with the degrees of freedom. Applying a LASSO penalty constraint to logistic models can greatly reduce over-fitting compared to the more popular standard (unconstrained) and stepwise logistic models. In general, the LASSO models reduced over-fitting by approximately 50%, a finding which was robust to the choice of response variable and the degrees of freedom. The LASSO models also produced the highest out-of-sample classification accuracy (as assessed by AUC).

The consequences of over-fitting were illustrated by comparing odds ratios between models differing by extent of over-fitting. The evidence points to inflated odds ratios and an increased likelihood of false-positive findings associated with models that over-fit. For instance, the magnitude of the treatment effect in modeling persistence fell by approximately one-fifth in the LASSO model compared to the standard and stepwise logistic models. Whilst the penalised regressions shrunk the magnitude of the estimated treatment effect for both response variables in this study, the substantive interpretation of the size of the effect remained largely unchanged. This should not detract from the fact that in some studies, penalised regressions will materially alter the central finding of a study from a significant to an insignificant treatment effect as compared with orthodox regression methods.

There is a natural and understandable reluctance to hold-back data for out-of-sample model evaluation, given the desire to use as much data as possible to produce robust estimates. The widely-held assumption seems to be that it is not possible to assess the extent to which a model may over-fit without holding-back data, so attempts to quantify over-fitting are conspicuous in the literature by their absence. This study shows that accurate estimates of out-of-sample model accuracy and hence over-fitting can readily be derived without holding-back data. The estimates of over-fitting are produced using tenfold cross-validation, a well-established method in statistical learning and machine learning. In the case of standard logistic regression, this simply involves estimating ten logistic regression models each on nine-tenths of the data and calculating the average AUC for the left out tenth. In small samples, robust estimates may involve increasing the number of folds up to a limit of N-1 (so-called leave-one-out cross-validation). The routine adoption of cross-validation to provide a powerful diagnostic on model over-fitting and out-of-sample accuracy is encouraged as standard practice.

The results clearly demonstrate that lower degrees of freedom are associated with greater over-fitting. In the models involving a ratio of training observations to candidate covariates of 36.4 (1,674/46), the amount of over-fitting in the logistic regressions without penalization was around 10%. This increased to approximately 30% when the degrees of freedom was reduced to 8.7 (400/46). Further evidence on the association between the degrees of freedom and amount of over-fitting for different settings would be helpful to researchers who may not have a good grasp of the circumstances in which over-fitting is likely to pose a problem.

Importantly, both penalized models and cross-validation are a standard part of most popular statistical packages. Thus, it is hoped that the results presented in this study will encourage research in this field to adopt these methods more widely and hence contribute to a more robust evidence base.

This study has several important limitations. First, the exact relationship between over-fitting and inflated odds ratios can only be summised in the absence of a simulation framework where the true associations are known. Simulation studies would be welcome to better understand this relationship and to formally quantify biases inherent in different modeling approaches under different settings. Second, on a related theme, the sensitivity analysis in this study was limited to two response variables and two levels of degrees of freedom, all on a single dataset. More comprehensive sensitivity analysis for different settings would help inform the generalisability of the key findings reported here. Third, this study applied just one method of penalization (the LASSO). Other methods, such as Elastic Net and Ridge, may have produced less over-fitting and higher out-of-sample accuracy. These solutions are also likely to return a different balance between the selected covariates and the magnitude of the coefficients. Fourth, conditional logistic regressions may be more appropriate than standard (unconstrained) and stepwise logistic regressions for the matched case-control sample design.28 Standard and stepwise logistic regressions were chosen for this study since they are the most widely adopted regression methods and the aim of this study was to appeal to as broad an audience as possible. Nonetheless, it is unlikely that the choice of method would affect the relative magnitude of coefficients and extent of overfitting which were the primary focus of this study.

# Conclusions

This study showed that:

* Penalized LASSO logistic regression models greatly reduced over-fitting compared to standard and stepwise logistic regressions.
* The LASSO models produced higher out-of-sample classification accuracy compared to the standard and stepwise models.
* The LASSO models were associated with odds ratios closer to unity, suggesting that standard and stepwise models were more likely to produce inflated odds ratios.
* Good estimates of over-fitting can be derived without holding-back data through use of cross-validation.
* Healthcare research stands to benefit from greater adoption of penalized regressions and cross-validation to estimate the extent to which a model may overfit.

# Transparency

## Declaration of Funding

# JR and MH both made substantial contributions to the conceptual design, interpretation of results and drafting of the manuscript.

## Declaration of Financial/Other Relationships

# The authors declare no competing interests or relationships.

# Acknowledgements

We would like to thank Ian Bonzani and Jon Korn for their insightful comments, Yan Xue and Hui Jin for programming assistance and Lee-Anne Beecher for copy editing.

# 

# References

3 Murphy KP. Machine Learning A Probabilistic Perspective. MIT press; 2012.

4 Tibshirani R. Regression Shrinkage and Selection via the Lasso. J R Stat Soc Ser A. 1996; 58:1.

5 van Loo HM, van den Heuvel ER, Schoevers RA et al. Sex dependent risk factors for mortality after myocardial infarction: individual patient data meta-analysis. BMC Medicine. 2014; 12:242.

6 Ayyagari R, Vekeman F, Lefebvre P et al. Pulse pressure and stroke risk: development and validation of a new stroke risk model Curr Med Res Opin. 2014; 30.12.

7 Walsh C, Hripcsak G. The effects of data sources, cohort selection, and outcome definition on a predictive model of risk of thirty-day hospital readmissions. J Biomed Inform. 2014; 52.

8 Siebeling L, Musoro JZ, Geskus RB et al. Prediction of COPD-specific health-related quality of life in primary care COPD patients: a prospective cohort study. NPJ Prim Care Respir Med. 2014; 24.

9 Halabi S, Lin C-Y, Kelly WK et al. Updated Prognostic Model for Predicting Overall Survival in First-Line Chemotherapy for Patients With Metastatic Castration-Resistant Prostate Cancer. Journal of Clinical Oncology. 2014;32(7):671-677.

10 Kohavi A. A study of cross-validation and bootstrap for accuracy estimation and model selection. IJCAI . 1995; 14:2.

11 Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983;70:1.

12 Austin PC, Grootendorst P, Anderson GM. A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study. Stat Med. 2007; 26:4.

13 Fawcett T. An Introduction to ROC analysis. Pattern Recognit Lett. 2006; 27:8.

14 Hesterberg T, Choi NH, Meier L et al. Least angle and ℓ1 penalized regression: A review. 2008;2.

15 James G, Witten D, Hastie T, Tibshirani R. An Introduction to Statistical Learning. Springer Texts in Statistics. 2013; 103.

16 Kuhn M, Johnson K. Applied Predictive Modelling. Springer; 2013. p. 101-137.

17 Varian HR. Big Data: New Tricks for Econometrics. J Econ Perspect. 2014; 28:2. doi: 10.1257/jep.28.2.3.

18 Zou H, Hastie T. Regularization and Variable Selection via the Elastic Net. . J R Stat Soc Ser B. 2005; 67:2.

19 Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal 24 problems. Technometrics. 1970;12:1. doi:10.1080/00401706.1970.10488634.

20 Lockhart R, Taylor J, Tibshirani R. A significance test for the Lasso. Annals of statistics. 2014;42(2):413-468

21 Clemmensen L, Hastie T, Witten D, Ersboll B. Sparse Discriminant Analysis. Technometrics. 2011; 53:4.

22 Witten D, Tibshirani R. Covariance–Regularized Regression and Classification For High Dimensional Problems. J R Stat Soc Ser B. 2009; 71:3.

23 Chun H, Keles S. Sparse Partial Least Squares Regression for Simultaneous Dimension Reduction and Variable Selection. J R Stat Soc Ser A: Series B. 2010; 72:1.

24 Jolliffe I, Trendafilov N, Uddin M. A Modified Principal Component Technique Based on the lasso. J Comput Graph Stat. 2003; 12:3.

25 Zou H, Hastie T, Tibshirani R. Sparse Principal Component Analysis. J Comput Graph Stat. 2004; 15:2006.

26 Kumar KT. Multicollinearity in Regression Analysis. Rev Econ Stat. 1975; 57:3

27 Piepho HP, Ogutu JO, Schulz-Streeck T et al. Efficient Computation of Ridge-Regression Best Linear Unbiased Prediction in Genomic Selection in Plant Breeding. Crop Sci. 2012; 52.

28 Gail MH, Lubin JH, Rubinstein LV  Likelihood calculations for matched case-control studies and survival studies with tied death times. Biometrika 68:703-707, 1980.































































# Supporting Information

**S1 Table. Odds ratios for 50/50 training/test samples (relapse) full results.**

**S2 Table. Odds ratios for 400 training/test samples (relapse) full results.**

**S3 Table. Odds ratios for 50/50 training/test samples (persistence) full results.**

**S4 Table. Odds ratios for 400 training/test samples (persistence) full results.**

**S5 Table. Odds ratios for logistic regression (relapse) on full sample (N=3348).**

**S6 Table. Odds ratios for logistic regression (persistence) on full sample (N=3348).**

**S7 Table. Odds ratios for stepwise logistic regression (relapse) on full sample (N=3348).**

**S8 Table. Odds ratios for stepwise logistic regression (persistence) on full sample (N=3348).**