Comparing the g-formula fitted with generalized linear models (GLMs) and g-formula fitted with machine learning methods (MLMs).

S4459997

26/06/2022

Supervisors – Prof. Dr. Eelko Hak and Dr. Maarten J. Bijslma

Unit PTEE

Introduction: One challenge that many variables in a dynamic treatment regime face is confounding, especially time-varying confounding. The g-formula is a method used to study dynamic treatment regimes. It helps in estimating the effectiveness of a model by using the counterfactual theory of causation. However, it is unclear how best to specify the g-formula.

Data and methods: To get insight into the problem of model specification in this study we fitted the g-formula with generalized linear models (GLMs) and machine learning methods (MLMs) and then compared them in terms of their predictions with the help of loss functions. Using a simulation study we try to determine the predictivity of each model from different classes. To visually distinguish the predictions of each model in this study namely, the generalized linear model (GLM), lasso model and the random forest model we make use of density plots.

Results: The density plots revealed that the generalized linear model (GLM) and the lasso predictions are very close to each other and are also quite similar to the validation dataset. The random forest plots were far off from the validation dataset. Further into the study the loss values obtained namely, the mean absolute error and the mean squared error indicated something similar to the density plots. Again the generalized linear model (GLM) and lasso model loss values were very close to each other while the random forest loss values were much higher than the other models.

Conclusion: Determining the predictivity is an important step towards determining the performance of the models in causal effect estimation. After looking at the density plots and the loss values, both the lasso model and the GLM model were recommended for future studies on causal effect estimation. Lasso model is a machine learning method and it might work better with large amounts of data. However we still do not know how well machine learning methods (MLMs) perform with causal effect estimation. Hence, GLM models are also recommended.

Layout

Section 1

Introduction

What is causal effect?

Suppose we compare two outcomes of which one includes an action A and the other not including the action A considering all other things remain equal. If the two outcomes differ, we can conclude that the action A has a causal effect (causative or preventive) on the outcome. In epidemiology action A is referred to as exposure or treatment [1].

In a hypothetical scenario, Travis is scheduled to undergo an appendectomy, Travis dies five days after the surgery. In some way we come to know that had Travis not undergone the surgery (provided all other things remain the same) he would be alive ten days later. Most people would surmise that the surgery caused the death of Travis. Thus, the appendectomy had a causal effect on Travis' ten-day survival. Alternately, another patient named Tina underwent an appendectomy on the same day but ten days later she was alive, now again somehow, we come to know that had Tina not undergone the surgery she would still be alive after ten days, Hence the surgery didn’t have a causal effect on Tina’s ten-day survival. These two examples represent how causal inference works. Here the link between causal effect and types of study is that from observation alone, making it difficult to determine causality. If only we had multiple Travis’ and expose some of them to the intervention and some of them remain unexposed to the intervention, we could compare the outcomes between the two groups and determine the effect of surgery on Travis’ survival. In the real world we don’t have multiple Travis’, so we have to come up with an alternative like, the RCT, which achieves this by randomizing the exposure to a group. In a simple trial one group of people are exposed to the new treatment while others remain unexposed, the results are then used to determine the causal effect. Experimental studies test the consequences of an intervention on the population as an example of a randomized test. Experimental studies such as Randomized clinical trials are considered as ‘gold standard’ for causality inferences. Alternatively, observational studies basically draw inferences from a sample to a population. In observational studies the treatment and the exposures are not under the control of the investigator whereas in experimental studies they are. Issues associated with ethics may arise concerning experimental studies also they're time-consuming and expensive hence, observational studies are often preferred. However, observational studies have a high risk of confounding.

What is confounding?

A confounding variable is a variable that impacts both the dependent and independent variables, which causes a spurious relationship I.e., where two or more variables are associated but not related causally. Alternately, confounding also takes place if there is a true causal relationship between the exposure and the outcome. In that case, the estimate of the true effect becomes larger or smaller.

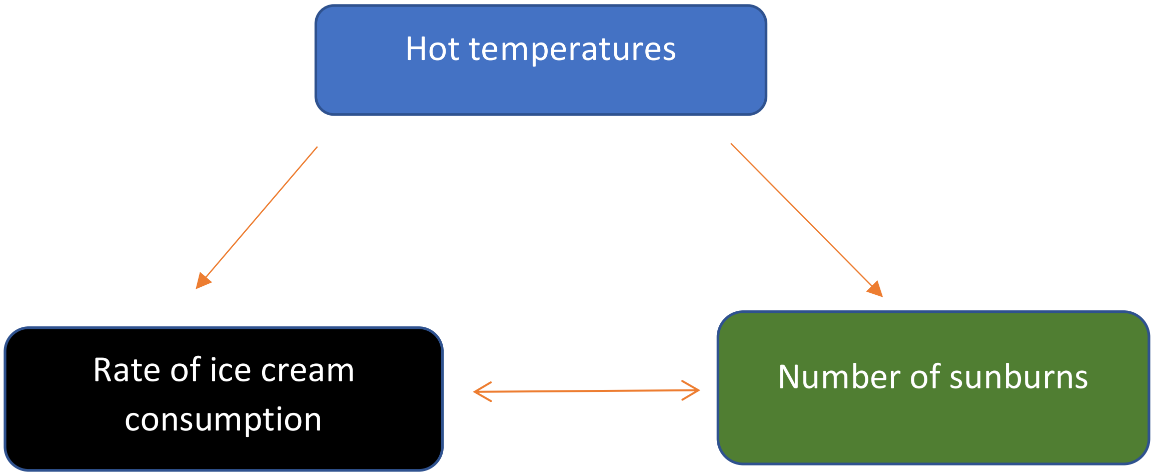

Figure.1, The relation of a confounding variable with the outcome, the confounding variable being 'hot temperatures'.

In figure.1, hot temperature is the confounding variable which means it causes both increase in rate of consumption and number of sunburns, but that does not mean that increase in rate of ice cream consumption leads to the outcome, which is increase in the number of sunburns.

There are different types of confounding such as time-constant confounding where the values of confounders do not change over time for example: sex, birthplace; time-varying confounding where the values of confounders change over time for example, body mass index (BMI), blood pressure, etc [2]. There are several methods to deal with confounding such as restriction, matching, randomization, stratification. The g-formula is method which is specifically devised to adjust for time-varying and intermediate confounders.

What is the G-formula?

The g-formula is one of the methods employed to study dynamic treatment regimes as in such regimes we often stumble across time-varying confounding. It's a technique developed by statisticians and used by both statisticians and epidemiologists which helps in estimating the effectiveness by using the counterfactual theory of causation I.e., the causal claim is elucidated with relevance to the counterfactual theory or conditions. According to *Keil et.al* "Unlike standard regression approaches the parametric g-formula can be used to adjust for time-varying confounders that are affected by prior exposures" [3] therefore, according to *Bijlsma et.al* "it's the potential to account for hypothetical changes during the method while also allowing for the interdependency between the determinants" [4]. Usually, the effectiveness is measured by either comparing two counterfactuals (everyone exposed vs nobody exposed) or by comparing the 'natural course' to the hypothetical intervention.

The natural course is an approximation of the empirical data which depicts the outcome if there is no intervention. Interventions are often measured by regression models and are adjusted for confounders, for instance a Cox regression model estimates specific hazard ratios and so averages it which could lead on to confounding away effects of the exposure. Generalized linear models (GLM) have risk of overfitting and a bias in time-dependent covariate which might be caused by a confounder and a causal mediator hence, g-formula is preferred. There are several modelling procedures which can be incorporated into the g-formula, commonly incorporated are GLMs, however, newer methods such as machine learning methods are also possible. Although, machine learning methods are at a greater risk of overfitting the data.

What is overfitting?

Overfitting is when a model fits too closely to the dataset it was developed with (training dataset) which could lead to wrong predictions and poor generalizations on testing data. This is because the model has high variance and low bias towards the training dataset. "Variance is how much a model changes in response to the training data" [5]. If the model predicts too closely to the training dataset then that means it has high variance. A model with high variance usually performs badly when applied to a dataset other than it was trained with (validation dataset). Precision is the converse of variance but if a model fits too closely to the data (it appears to have high precision) then it results in a bias. A low bias may seem like a beneficial thing but each model makes assumptions about the data and should leave room for differences which cannot be seen in the training dataset.

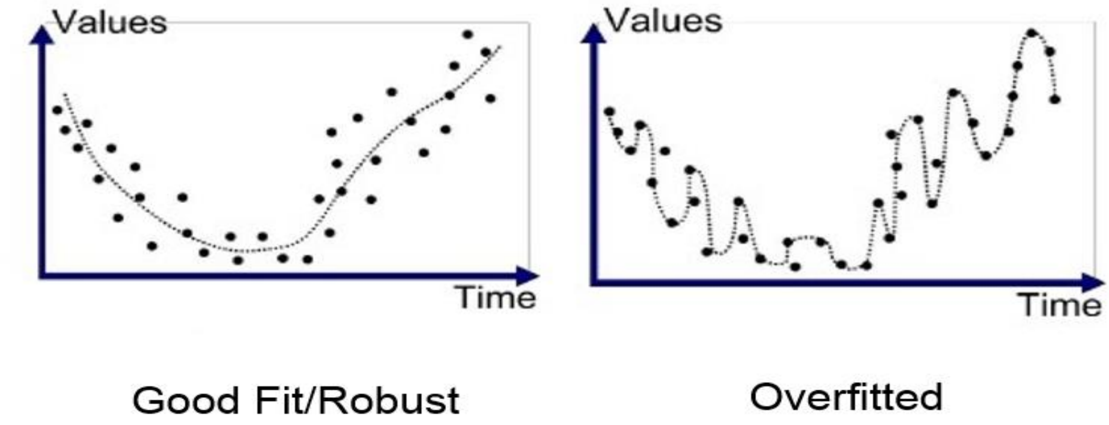

Figure 2. An example of overfitting.

In figure 2, the points in the graph represent the data and the dotted lines represent the model predictions. The graph on the right depicts the model which overfits the data as we can that it predicts too close to the training dataset. Whereas the graph on the left represents a good fit or a robust model.

One of the methods to deal with overfitting is cross-validation. According to Brownlee "Cross-validation is a statistical method used to estimate the skill of machine learning models. It is commonly used in applied machine learning to compare and select a model for a given predictive modeling problem because it is easy to understand, easy to implement, and results in skill estimates that generally have a lower bias than other methods" [6]. Some of the types of cross validation are k-fold cross validation, leave-one-out cross validation, etc.

The aim of this thesis is to compare generalized linear models to machine learning methods such as regression trees inside a g-formula and see which fits better to the validation dataset.

Research question

To what extent does a g-formula fitted with underlying machine learning methods (regression trees, lasso) and a g-formula fitted with underlying generalized linear model fit to complex data from a dynamic treatment regime?

Section 2

Data and Methods

As the aim of this thesis is to compare g-formula fitted with generalized linear models (GLM) and g-formula fitted with machine learning methods (MLM), it is important to know the basics of these two methods. GLMs are interpreted by two components one that the distribution of the dependent variable should be a member of the exponential family which includes probability distributions like normal, binomial, Poisson distributions. Secondly the link function basically converts a non-linear relationship to a linear one so that the linear model can be a fit. By doing so it describes how the mean of the outcome and the linear combination of the predictors are related [7]. In contrast to GLMs, Machine learning methods such as regression trees make use of algorithms in their construction. Regression trees are decision trees (predictive models) where the dependent variable can take categorical as well as continuous values [8]. The two modelling approaches will be compared through cross validation to see which fits better to the data. Data will be simulated through the data generating process, as described below.

Data-generating process

There are many processes to generate data where it can be generated from already studied models by producing parametric draws or by repeated resampling from a dataset with replacement [9]. However, the data generating process in this thesis consists of pure simulations I.e., not based on real data. Pure simulations allow the coefficients in the data simulation to be known and can be used to compare the estimates. Multiple scenarios will be investigated going from simple to more complicated. A scenario was simulated for asthma treatment which is visually represented in Figure 3.

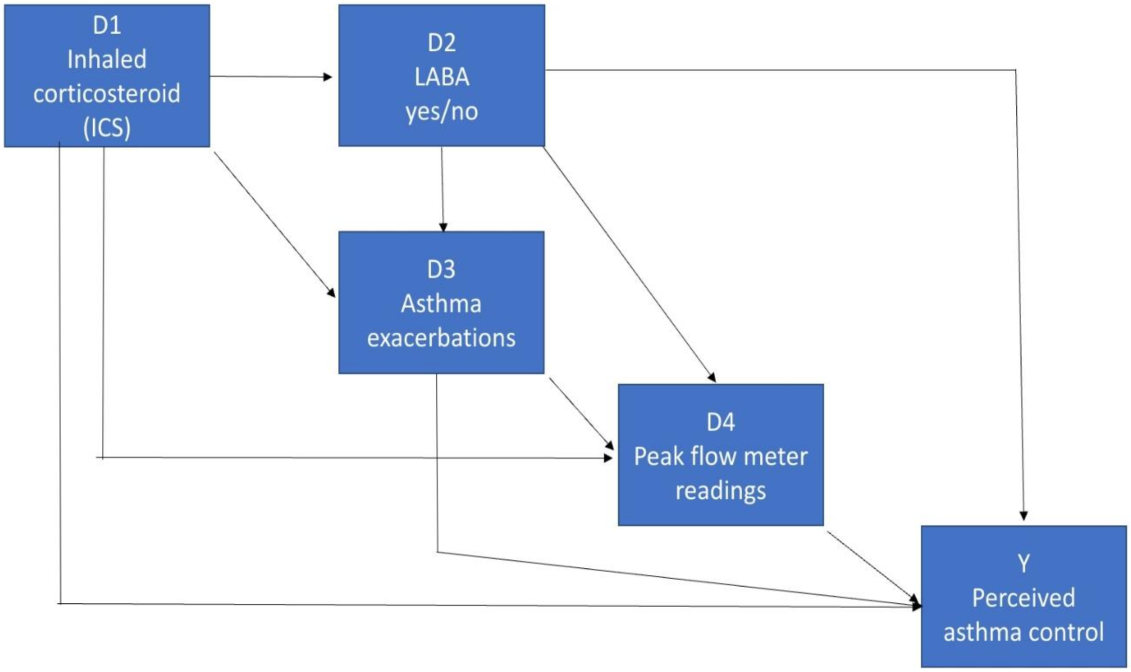

Figure 3. Flow diagram of the factors influencing each other

In figure 3, The D1 box represents the people who take different dosage levels of Inhaled corticosteroids (ICS), D2 box represents a binomial variable where the population either takes long-acting beta agonists (LABA) or does not. D3 box represents the number of exacerbations a person has. D4 box represents the peak flow meter readings of a person. The Y box represents the outcome variable which is the perceived asthma control which is measured by the visual analogue scale (VAS). All of the variables affect the outcome variable Y.

Below is the equation we use to generate the data.

$$\delta_1 \sim Unif(0,5)$$

$$\delta_2 \sim B\,(-0.5 + \delta_1 . 0.2)$$

$$\delta_3 \sim Pois(1 + \delta_1 + \delta_2 + \delta_1 . \delta_2 . 0.1)$$

$$\delta_4 \sim N(100(5 + \delta_1 + \delta_2 + \delta_3 + (\delta_1 . \delta_3) . 0.01) . 20{,}50)$$

$$Y \sim N(10 + \delta_1 . 1 + \delta_2 . 1 + \delta_3 . 1 + \delta_4 . 1{,}5)$$

Where Unif refers to uniform distribution N refers to normal distribution, Pois refers to Poisson distribution and B refers to binomial distribution, we take as our outcome of interest.

Estimation

Parameterization

Once the data is generated we use the following equations in the GLM and Lasso to estimate relationships in the data, as if they were empirical data to be analyzed

$$\delta_2 \sim B\left(\beta_{2,0} + d_1\beta_{2,1}\right)$$

$$\delta_3 \sim Pois\left(\beta_{3,0} + d_1\beta_{3,1} + d_2\beta_{3,2}\right)$$

$$\delta_4 \sim N\left(\beta_{4,0} + d_1\beta_{4,1} + d_2\beta_{4,2} + d_3\beta_{4,3}\right)$$

$$Y \sim N\left(\beta_{5,0} + d_1\beta_{5,1} + d_2\beta_{5,2} + d_3\beta_{5,3} + d_4\beta_{5,4}\right)$$

Where β refers to the unknown coefficient values.

Monte Carlo (MC) integration:

Following the estimation of the earlier specified models, we then use Monte Carlo (MC) integration to estimate quantities in our g-formula. In the GLM and Lasso we simulate by taking random draws from the following specified distributions:

$$\tilde{d}_2 \sim B(d_1\hat{\beta}_{2,1})$$

$$\tilde{d}_3 \sim Pois(d_1\hat{\beta}_{3,1} + \tilde{d}_2\hat{\beta}_{3,2})$$

$$\tilde{d}_4 \sim N(d_1\hat{\beta}_{4,1} + \tilde{d}_2\hat{\beta}_{4,2} + \tilde{d}_3\hat{\beta}_{4,3} + \hat{\sigma}_4^2)$$

$$\tilde{Y} \sim N(d_1\hat{\beta}_{5,1} + \tilde{d}_2\hat{\beta}_{5,2} + \tilde{d}_3\hat{\beta}_{5,3} + \tilde{d}_4\hat{\beta}_{5,4} + \hat{\sigma}_5^2)$$

Where $\hat{\beta}$ refers to estimated coefficients from our earlier specified models, $\tilde{d}$ and $\tilde{Y}$ refers to the simulated values and $\hat{\sigma}^2$ refers to the variance of the residuals.

Fitting the models

The data was simulated through the data generating process. The simulated data was separated into two sets namely the training dataset and a validation dataset, where the training dataset i.e., 70% of data is used to fit the model on and then predict the remaining 30% data which is called the validation dataset.

The generalized linear models and the Lasso models will be fit following the earlier specified parametrizations. In case of MLM, more specifically the regression tree, we do not tell the model the exact specification but we do tell it which terms to include i.e., once the model knows which variables to include it will figure out rest of the functional form itself [10]. Although other machine learning methods like LASSO require separate fitting just like GLMs [11].

To compare which modelling approach works best, the loss for each model was calculated. The loss in the models will be calculated by loss functions. Loss functions are measures of how well set the parameters are, basically it quantifies the difference between the current output of the algorithm with the expected output [12]. In this thesis we will use two loss functions, sum of squared errors (SSE) and sum of absolute errors (SAE). SSE as the name suggests is the sum of the squares of the residuals which are the deviations from the empirical data the models predict. Lesser the value of SSE the better a fit the model is considered. SAE is the sum of the absolute difference between the predicted value and the ‘true’ value.

$$SSE = \frac{1}{n_{val}} \sum_i (y_i - f_i)^2$$

$$SAE = \frac{1}{n_{val}} \sum_i |(y_i - f_i)|$$

Where y_i is the observed data value, f_i is the predicted value, a value near 0 indicates that there is a small chance of error and that the fit is better at predicting values.[13]

To stabilize the findings, we performed the loss function for each Monte Carlo iteration (i.e. 50 iterations) and took the average over the calculated loss values.

Results

Below are the density plots of each variable used in this treatment, represent the predictions of each model when compared to the validation data. Density plots are often used to determine the distribution of a variable in a given dataset.

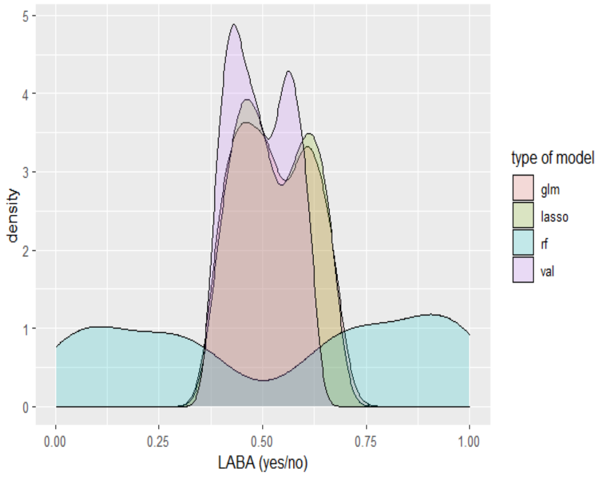

Figure 4. Density plot of LABA

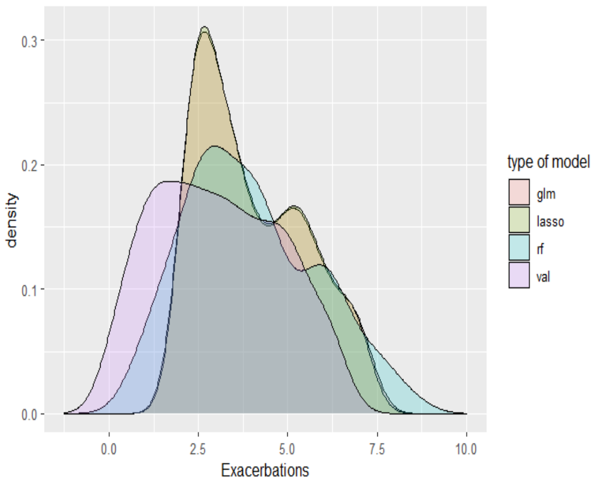

Figure 5. Density plot of Exacerbations

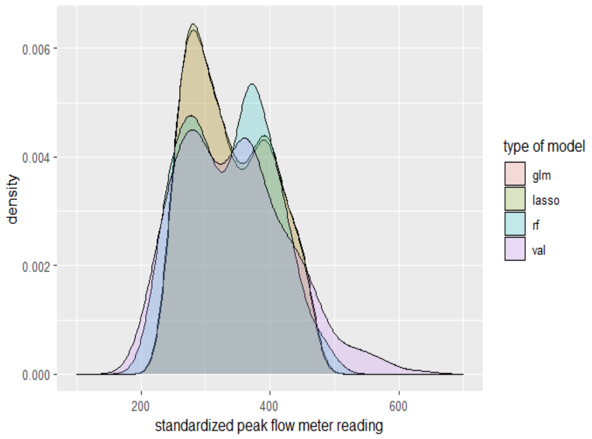

Figure 6. Density plot of peak flow meter reading

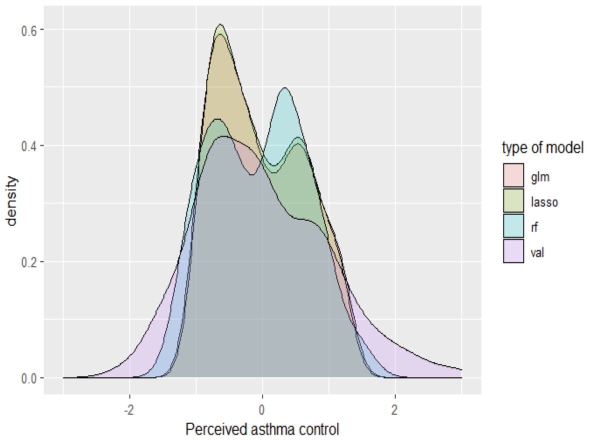

Figure 7. Density plot of Perceived asthma control

In figure 4, we can see that the curves for GLM and Lasso are very close to each other but the Lasso is slightly closer to the validation curve than the GLM curve.

The random forest curve performs badly. As we can see that the Random forest plot has two prominent peaks, this is because the random forest in principle classifies and divides the data.

In figure 5, The GLM and Lasso curves are similar again. The random forest curve seems to be more similar to the validation curve in terms of the distribution of the curve i.e., the random forest curve is the closest in depicting the spread or the variance of the validation data when compared to Lasso and GLM curves.

In figure 6, The GLM and Lasso curves predict the peaks better than the random forest as it puts too much weight on one of the peaks which could lead to wrong predictions. The random forest model is the only one which comes close in depicting the spread of the validation dataset.

In figure 7, the curves are similar to figure 6.

In all the figures we can notice two prominent peaks for all the model types, this is because each model prediction is based on the previous figure i.e., figure 5 predictions are based on figure 4 predictions.

	Random Forest	**Lasso**	**GLM**
d2	163.02	163.25	162.80
d3	623.48	561.18	561.25
d4	19856.15	18598.96	18591.54
Y	208.4226	195.09	195.00

Table 1. Mean absolute loss values (the purple boxes represent the lowest loss value among the three models)

	Random Forest	**Lasso**	**GLM**
d2	119.60	82.03	81.99
d3	1935.06	1542.96	1543.56
d4	1884592	1645233	1645504
Y	207.74	180.89	180.92

Table 2. Mean squared loss values (the purple boxes represent the lowest loss value among the three models)

The Monte Carlo error reduction for each model was run fifty times and then took the mean of each loss value to avoid any chance result. The tables above depict the mean loss values, wherein the pink boxes represent the lowest loss value of the three models. As we can see that in table 1 for mean absolute loss the GLM model predicts better than the rest for all variables except one. In table 2 for the mean squared loss values the Lasso model predicts better than the rest for all variables except one. The random forest model as we can see in the tables has much higher loss values in comparison to the Lasso and GLM models. Both the GLM and Lasso model predictions are very close to each other hence as far as predictions are concerned both GLM and Lasso are better at predicting the data. However, we can notice that GLM has slightly lower mean absolute values when compared to Lasso, while the Lasso has slightly lower mean squared loss values when compared to the GLM.

Discussion

The aim of this study was to compare generalized linear models (GLM) and machine learning methods (MLM) incorporated in the g-formula using cross-validation techniques. We employed the g-formula method in an asthma treatment regime wherein the variables in the treatment regime influence each other.

After the results were obtained, we found that the GLM and Lasso models predictions are very close to each other whereas the random forest model predictions perform worse. This can be observed from the loss values which were obtained after running the Monte Carlo error reduction fifty times, the density plots provide a visual representation of the predictions of each model when compared to the validation data and they showed that the Lasso and GLM models are close in depicting the peaks of the validation data while the random forest is close in depicting the spread of the validation data.

The result clearly displays the failure of the random forest to predict data, this is because random forest deals best with data that is more spread out as it has much greater variance and in principle it divides the data into several parts which can be seen in figure 4 for a better understanding. The random forest that we use here was a simple one hence it could only bifurcate data broadly, a more complicated and detailed random forest would predict better with this data. The Lasso performs better with mean squared loss values because it 'shrinks' the coefficients towards zero this in turn reduces the variance and hence the data can be said to be condensed, this reduction in variance is the main cause for reduced mean square loss values. As the coefficients of GLM are more varied in comparison to Lasso the square function penalizes the outliers which causes the values of the GLM to be higher than Lasso. The mean absolute loss does not have the square function and hence variance in the model does not matter. Thus, the GLM performs better than the Lasso in terms of mean absolute loss values.

Determining the predictivity of the models is the first step in providing a pathway for future research on the causal effects. The next step of future research on the topic should be to compare the ability of the three models namely the GLM, Lasso and Random forest models in determining the causal effects. We know that the GLM models perform well with causal effect estimation in a g-formula, but we do not know how well MLMs do with causal effect estimation in a g-formula. The recommendations for future research based on this study would be to test more machine learning methods such as neural networks, Bayesian networks and support vector machines (SVM). Neural networks for instance are useful in modelling non-linear processes. Whereas Bayesian networks have a directed acyclic graph (DAG) in their structure which allows them to depict causal relations between variables [14]. SVMs are linear classifiers while neural networks and random forest are non-linear classifiers, this can provide a different perspective for machine learning methods. In future studies also simulating the lasso and the random forest with more covariates is recommended as patient who report at time point 0 for example in this treatment regime already have a large medical history. Hence, taking that into account would provide a better insight in the causal effect estimation. Also, MLMs might perform more accurately than GLM when there is a lot more data [15] [16].

In this study we noticed that the GLM and Lasso model predictions were very close in to each other. The random forest predictions were not close to the validation dataset. However, if the dataset is more spread out then the random forest model might predict better than the rest of the models. The recommendation generally would be to use Lasso and GLM models for future research, although Lasso is more preferred to GLM models because MLMs might perform better than GLM when there is large data. Hence, in a real world scenario when there is a lot more data at hand, Lasso models would be more effective in predicting data.

References

[1]. Hernán MAA definition of causal effect for epidemiological research Journal of Epidemiology & Community Health 2004;58:265-271.

[2] Mansournia et.al. , Handling time varying confounding in observational research; BMJ 2017; 359 doi: https://doi.org/10.1136/bmj.j4587 (Published 16 October 2017).

[3] Keil AP, Edwards JK, Richardson DB, Naimi AI, Cole SR. The parametric g-formula for time-to-event data: intuition and a worked example. Epidemiology. 2014;25(6):889-897. doi:10.1097/EDE.0000000000000160.

[4] Bijlsma et.al. Modelling the socio-economic determinants of fertility: a mediation analysis using the parametric g-formula Max Planck Institute for Demographic Research, Rostock, Germany [Received December 2017. Final revision August 2019 https://doi.org/10.1111/rssa.12520.
[5] https://towardsdatascience.com/overfitting-vs-underfitting-a-complete-example-d05dd7e19765
[6] Jason Brownlee A Gentle Introduction to k-fold Cross-Validation machinelearningmastery.com May 2018.
[7] Faraway, J. J. (2006). Extending linear models with R. Boca Raton, Fla: Chapman & Hall/CRC pg 139.

[8] Wu, X., Kumar, V., Ross Quinlan, J. et al. Top 10 algorithms in data mining. Knowl Inf Syst 14, 1–37 (2008). https://doi.org/10.1007/s10115-007-0114-2

[9] Morris, TP, White, IR, Crowther, MJ. Using simulation studies to evaluate statistical methods. Statistics in Medicine. 2019; 38: 2074– 2102. https://doi.org/10.1002/sim.8086

[10] Pere, Christoph June 2020 https://towardsdatascience.com/what-is-loss-function-1e2605aeb904

[11] Tony Yiu, Understanding Random Forest, June 2019

https://towardsdatascience.com/understanding-random-forest-58381e0602d2

[12] Musoro, Jammbe Z et al. "Validation of prediction models based on lasso regression with multiply imputed data." *BMC medical research methodology* vol. 14 116. 16 Oct. 2014, doi:10.1186/1471-2288-14-116

[13] https://web.maths.unsw.edu.au/~adelle/Garvan/Assays/GoodnessOfFit.html

[14] Pekka Parviainen, Bayesian Networks, University of Bergen. October 2019

https://www.uib.no/en/rg/ml/119695/bayesian-networks

[15} Sunil Ray, 8 Proven Ways for improving the "Accuracy" of a Machine Learning Model. December 2015

https://www.analyticsvidhya.com/blog/2015/12/improve-machine-learning-results/

[16] Watt J, Borhani R, Katsaggelos AK. Machine Learning Refined: Foundations, Algorithms, and Applications. 2nd ed. Cambridge: Cambridge University Press; 2020.