MarketScan

Accelerate research and discovery with machine learning in HEOR and epidemiology studies

A comparison of machine learning methods in health economics and outcomes research and epidemiology

merative

Introduction

For decades, researchers have used traditional regression methods to answer pressing questions for life sciences and healthcare research. However, with the increasing capabilities of machine learning (ML) methodologies, researchers have adopted more complex algorithms to generate compelling insights using real-world data. Proper application of machine learning methodologies can help life sciences organizations drive their research agendas forward and develop solid evidence for communication to key stakeholders such as regulatory bodies, payers, and providers.

Use of machine learning, however, depends not only on a researcher's ability to perform the analyses, but also on the researcher's nuanced understanding of project goals, as well as advantages and disadvantages of various methods. For example, ML methodologies may have varying degrees of predictive capability, statistical inference, and interpretability, and may require differing amounts of computing resources. Furthermore, model selection can depend on the specific predictor and outcome variables included in the analysis.

This case study, written by Merative® outcomes research experts, compares five supervised learning algorithms and their considerations for use in the context of health economics and outcomes research. This case study demonstrates that there are multiple factors involved in model selection, and a strong understanding of these factors, as well as the advantages and disadvantages of each model type, are necessary when designing and conducting HEOR and epidemiology studies.

About Merative researchers

Merative researchers have decades of experience conducting outcomes research, consulting, and collaborating on the execution of pre- and post-launch health economics and outcomes research agendas. Merative experts conduct research across a range of therapeutic areas and regularly employ a variety of advanced methodologies such as those demonstrated in this paper, backed by robust proprietary and public data assets.

Case study objective

The objective of this case study is to predict which newly-diagnosed multiple sclerosis (MS) patients are at risk for an inpatient stay within a year of diagnosis.

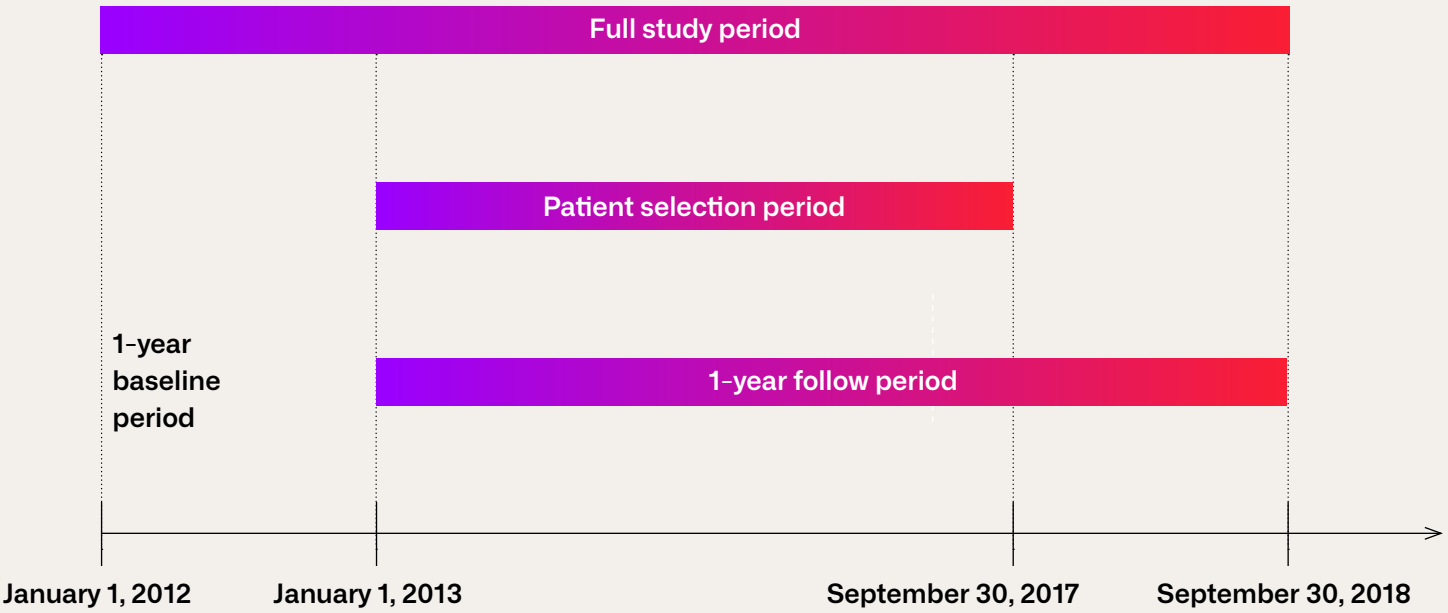
Study design

Merative researchers used the Merative MarketScan® Commercial and Medicare Supplemental Database to complete this analysis. MS was defined as having ≥1 inpatient claim with a primary diagnosis of MS, having ≥2 non-diagnostic outpatient claims, or having ≥1 non-diagnostic outpatient claim with a diagnosis of MS and ≥1 fill for a disease modifying agent. Index date was defined as the time of first claim (inpatient or outpatient) with a diagnosis of MS. The patient selection period occurred between January 1, 2013 and September 30, 2017.

Patient selection criteria

Criteria	N
Diagnosis of MS from 1/1/2012 through 4/30/2018 (first diagnosis serves as the index date)	152,892
Aged ≥18 years at index	152,178
Continuous medical and pharmacy eligibility for at least 12 months prior to index	27,515
No evidence of MS diagnoses or DMDs in the pre-period	26,513
Continuous medical and pharmacy eligibility for at least 12 months following index	18,381

Study timeline



Variables of interest

The outcome of interest is the presence of any inpatient (IP) visit within 1 year after MS diagnosis.

Predictors include the following:

Demographics

Baseline Healthcare Utilization and Costs

Charlson Comorbidity Index

Number of MS Diagnoses

Pre-Index Medications
- Anticonvulsants
- Antidepressants
- Antibiotics
- Antifungals
- Corticosteroids
- Immunosuppressants
- Muscle Relaxants
- Stimulants
- Opioids
- NSAIDs

Pre-Index Comorbidities
- Depression
- Anxiety
- Arthritis
- Diabetes
- Fatigue
- Hyperlipidemia
- Hypothyroidism

Kurtzke's Functional Symptoms
- Pyramidal
- Cerebellar
- Brainstem
- Sensory
- Bowel and Bladder
- Mental
- Optic

About the data

Merative MarketScan Commercial Database consists of medical and drug data from employers and health plans for over 293 million individuals annually, encompassing employees, their spouses and dependents who are covered by employer-sponsored private health insurance in the US.
Merative MarketScan Medicare Supplemental Database includes the Medicare-covered portion of payment (represented as Coordination of Benefits Amount or COB), the employer-paid portion and out-of-pocket patient expenses.

Methods

After randomly splitting the data into a training dataset containing 75% of patients and a testing dataset containing 25% of patients, researchers used five different algorithms to estimate the probability of any inpatient visits. The algorithms used were multivariable logistic regression, elastic net regression, a decision tree, a random forest, and a neural network. The training dataset was randomly down-sampled to achieve a 1:1 ratio of patients with IP visits to controls because decision trees, random forests, and neural networks are sensitive to class imbalances. Tuning parameters were selected using 10-fold cross validation repeated 10 times. While many more machine learning algorithms exist, this white paper will focus on five supervised learning algorithms commonly used in HEOR and epidemiology studies. Each algorithm has its own strengths and limitations which should be taken into consideration when deciding which models to use in an analysis. See Appendix A for a description of each algorithm and the advantages and disadvantages of each. Predictive performance of machine learning models can be calculated by a variety of statistics. For the purposes of this case study, model performance was presented using the C statistic and Brier score, as is common with classification models. Please see Appendix B for more information on these performance measurements.

Results

Logistic regression

This model predicts any MS-related IP stay within a year of MS diagnosis as a function of predictors.

Pre-index ER utilization, type 2 diabetes, hypertension, and anticonvulsant prescriptions were associated with significantly higher odds of an IP stay within a year after MS diagnosis. Increasing age, any pre-index neurologist visit, and increasing number of pre-index office visits were associated with significantly lower odds of having an inpatient admission within a year of MS diagnosis.

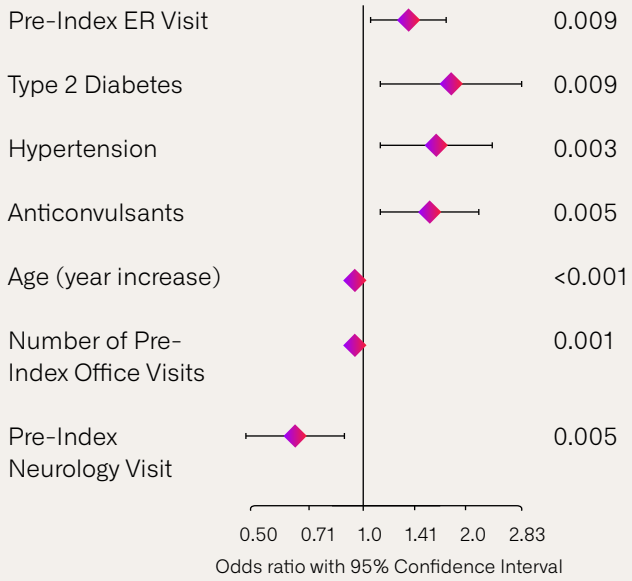
C Statistic

0.675

Brier Score

0.0218

Important predictors



Elastic net regression

Like the logistic regression model, the elastic net regression model predicts any MS-related IP stay within a year of MS diagnosis as a function of predictors. The most important predictors include age, presence of and number of ER visits, whether the patient was a child or spouse of the plan holder, immunosuppressant or biologic use, sensory symptoms, pre-index healthcare costs, number of office visits, and antidepressant use.

Notice that the coefficients are much smaller than those from logistic regression because elastic net models shrink coefficients towards zero.

C Statistic

0.683

Brier Score

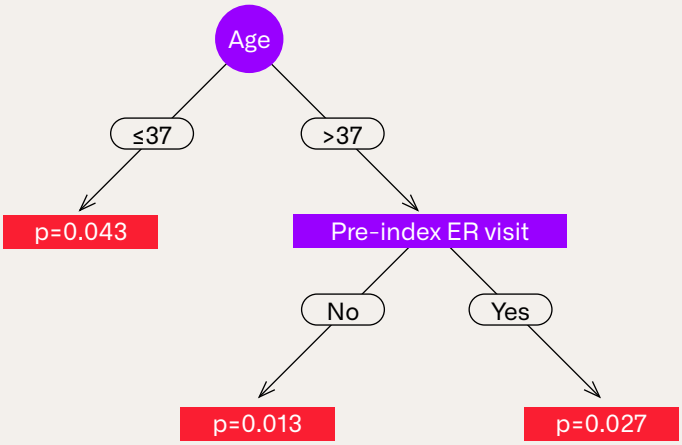
0.0218

Top 10 predictors using elastic net regression

Predictor	Odds ratio
Age	0.954
Number of ER Visits	1.031
Child of Plan Holder	1.031
Any ER Visit	1.031
Immunosuppressant/Biologic Use	0.979
Spouse of Plan Holder	0.979
Sensory Symptoms	1.020
Pre-Index Healthcare Costs	0.982
Number of Office Visits	0.983
Antidepressant Use	0.986

Decision tree

A decision tree was produced using the conditional inference tree algorithm, predicting any IP stay within a year after MS diagnosis using MS predictors. The complexity of the tree was determined by selecting a tuning parameter using repeated cross validation. The figure illustrates the tree and predicted probability of having an MS-related IP visit within each subgroup. Younger patients had the highest probability of having an MS-related IP-visit. Among patients older than 37, those with any pre-index ER utilization had a higher risk of a post-index inpatient admission.



C Statistic

0.638

Brier Score

0.0218

Random forest

The random forest model predicts the probability of any IP stay within a year after MS diagnosis using all MS predictors. This model does not produce odds ratios because it aggregates predictions from hundreds of individual decision trees. Instead, we can estimate each predictor's importance on the final predictions generated. Age, pre-index healthcare costs, number of various categories of outpatient visits, any ER visit, sensory symptom, being a spouse of the plan holder, and any neurology visit are most important for generating the final predictions.coefficients towards zero.

C Statistic

0.687

Brier Score

0.0218

Top 10 predictors using random forest model

Predictor	Odds ratio
Age	11.26
Pre-Index Healthcare Costs	8.78
Number of Other Outpatient Visits	7.20
Number of Office Visits	6.89
Number of ER Visits	6.42
Number of Neurology Office Visits	4.42
Any ER Visit	4.37
Sensory Symptoms	2.90
Spouse of Plan Holder	2.59
Any Neurology Office Visit	2.55

Neural nework

Neural network predicts probability of any IP stay within a year after MS diagnosis using all MS predictors.

Neural networks do not produce easily interpretable odds ratios or figures, so Garson's variable importance was estimated for each variable.

The most important predictors include presence of an ER visit, number of ER visits, age, immunosuppressant or biologic use, sensory symptoms, any neurology office visits, being a child or spouse of the plan holder, pre-index neurology costs, and urban versus rural location.

C Statistic

0.687

Brier Score

0.0218

Top 10 predictors using neural network model

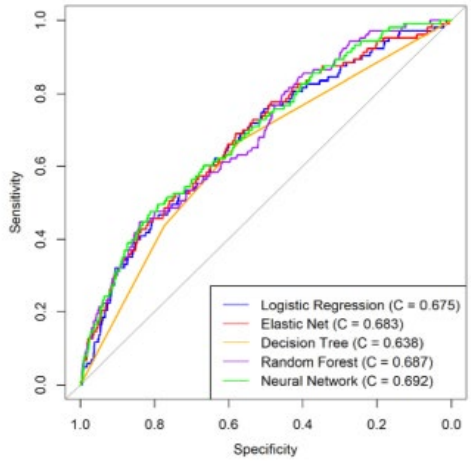
Predictor	Importance
Any ER Visit	0.066
Age	0.065
Immunosuppressant/Biologic Use	0.047
Sensory Symptoms	0.047
Number of ER Visits	0.045
Any Neurology Office Visit	0.039
Child of Plan Holder	0.035
Spouse of Plan Holder	0.033
Pre-Index Neurology Costs	0.029
Urban vs. Rural Location	0.025

Comparison of results

The neural network performed best out of the five models. The neural network, random forest, and elastic net machine learning models modestly outperformed traditional logistic regression. However, differences in model performance were small.

The receiver operator curve, which shows the sensitivity and specificity of models, can be used to visualize and compare models. The best models aim to maximize both sensitivity and specificity.

Model	ROC-AUC	Brier score
Logistic Regression	0.675	0.0218
Elastic Net Regression	0.683	0.0218
Decision Tree	0.638	0.0218
Random Forest	0.687	0.0217
Neural Net	0.692	0.0217



Comparison of Receiver Operator Curves

Summary

Machine learning models discriminated whether patients will have an inpatient visit within a year of MS diagnosis better than logistic regression, though performance differences were small.

Model selection requires a thorough understanding of the benefits and drawbacks of various algorithms in the context of the project at hand. Regression and decision trees are useful in situations where statistical inference or highly interpretable models are desired, whereas random forests and decision trees can produce more accurate predictions when predictors have a complex relationship with the outcome of interest at the cost of interpretability.

It is important to remember that no modeling approach will work well if the training data do not contain key predictors of the outcome. Robust sources of real-world data combined with a rigorous and informed analytical approach can support appropriate model selection for HEOR and epidemiological research.

Conclusion

Machine learning can enhance HEOR and epidemiology studies by providing fit-for-purpose analytics when applied appropriately. While in many circumstances machine learning methodologies can provide analytical rigor to drive more discerning results, it is important to assess project goals to determine the best model. Working with a trusted partner that understands the nuances of different methodologies and when to apply various models can help ensure that the most appropriate algorithms are used. Furthermore, use of high-quality inputs in the form of in-depth, longitudinal real-world data is also an important component of developing reliable models.

To learn more about how Merative Life Sciences researchers and data scientists can support your organization's value demonstration strategy using robust real-world data, advanced methodologies, and complex study design, please contact your Merative representative or reach out via https://merative.com/contact.

Appendix

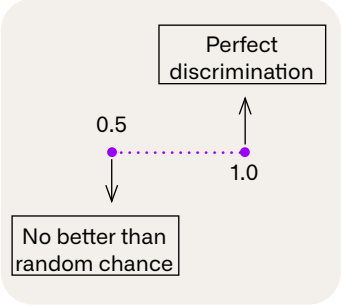
Appendix A. Description of common maching learning methods in HEOR and their advantages and disadvantages

Method	Description	Advantages	Disdvantages
Regression	Predicts the outcome of interest by estimating the best linear relationship between predictors and the outcome	– Simple interpretation and statistical inference – Fast run times	– Requires a linear relationship between predictors and the outcome
Elastic Net Regression	Regression model that incorporates a small amount of bias to improve precision and/or removes irrelevant predictors to improve predictive ability	– Improved predictive ability compared to traditional regression – Fast run times	– Requires a linear relationship between predictors and the outcome – Less interpretable than traditional regression
Decision Trees	Divides data into distinct subgroups based on a series of if/then questions, each with unique predictions	– Models non-linear relationships between predictors and the outcome	– Assumes all subjects in a subgroup behave identically
Random Forests	Constructs many decision trees (a forest) from random subsets of the data and outputs the mean prediction of all trees	– Models non-linear relationships – Outperforms both regression and decision trees in many situations	– More complicated interpretation than regression or decision trees – More computationally expensive
Neural Nets	Designed to mimic human neuron connections. Uses layers of data transformations to extract features from data	– Flexible approach that often outperforms simpler algorithms – Can be built based on structured or unstructured data (e.g., images)	– Difficult to interpret why predictions are generated – Computationally expensive

Appendix B. Definition and interpretation of C-statistic and Brier score

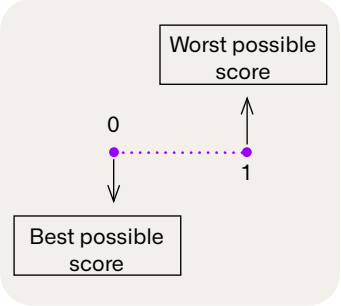
C Statistic (Concordance Statistic):

– Area under the receiver operator curve
– Measures a model's ability to discriminate the outcome
– Ranges from 0.5 to 1.0, where 0.5 indicates that the model is no better than random chance and 1.0 indicates perfect discrimination



Brier Score:

– Squared error of a probabilistic model; measure of model accuracy
– Range: 0 (best possible score) to 1 (worst possible score)



Other performance metrics for classification models include:

– Sensitivity
– Specificity
– Positive predictive value
– Negative predictive value
– Accuracy

About MarketScan

MarketScan by Merative provides deidentified, longitudinal, patient-level closed claims and specialty data for 293M+ patients sourced directly from a diverse pool of payers. Industry-leading researchers rely on MarketScan to derive valuable insights pertaining to health economics and outcomes research, treatment patterns, and disease progression across the industry resulting in more than 3,500 peer-reviewed manuscripts.

Learn more at
merative.com/real-world-evidence

About Merative

Merative is a data, analytics and technology partner for the health industry, including providers, payers, life sciences companies and governments. With trusted technology and human expertise, Merative works with clients to drive real progress. Merative helps clients reassemble information and insights around the people they serve to improve healthcare delivery, decision making and performance. Merative, formerly IBM Watson Health, became a new standalone company as part of Francisco Partners in 2022.

Learn more at merative.com

References

Kuhn, M., and Johnson, K. (2013). Applied predictive modeling. New York: Springer.

Merative
100 Phoenix Drive
Ann Arbor, MI 48108

Produced in the United States of America, June 2022.

MSN-3078997560 Rev 3.0