Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Comparing Variable Selection Methods in a Classification Framework: A Simulation Study

Samuele Minari

s.minari@campus.unimib.it

University of Milano-Bicocca

Antonella Zambon

University of Milano-Bicocca

Dario Pescini

University of Milano-Bicocca

Mirko Cesarini

University of Milano-Bicocca

Davide Soranna

IRCCS Istituto Auxologico Italiano

Research Article

Keywords: Variable selection, Simulation, Logistic Regression, Classification, Map Reduce

Posted Date: November 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11433379/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Variable selection is a critical step in clinical research, where datasets often contain many, and sometimes highly correlated, predictors. The goal is to identify the most relevant variables to improve model interpretability, parsimony, and predictive accuracy. However, data-driven selection methods can introduce substantial risks, including model instability, biased estimates, and the misidentification of relevant features. Following the ADEMP framework, we designed a comprehensive simulation study, that varied sample size, number of predictors, number of true signals, signal strength and correlation, to neutrally compare the performance of five variable selection methods in binary logistic regression across a range of simulated clinical-like scenarios. We evaluated five variable selection methods: stepwise logistic regression, LASSO, Elastic Net, Random Forest with out-of-bag (OOB) error-based backward elimination, and a Genetic Algorithm. Each method was assessed using 1,000 Monte Carlo simulations per scenario. Performance metrics included Area Under the Curve (AUC), True Positive Rate (TPR), False Discovery Rate (FDR) and the number of variables selected. The detailed results of these evaluations are presented, and the methods' performance is discussed across simulation scenarios.

1. Introduction

Variable selection is a key step when modeling the association between an outcome and several independent predictors. Its primary goal is to identify meaningful and statistically significant predictors while excluding irrelevant or redundant ones, resulting in more parsimonious and interpretable models[1–3].

Despite these advantages, it can introduce unintended challenges that are frequently underestimated by researchers. It may increase uncertainty in parameter estimation, introduce bias in regression coefficients, and lead to model instability (i.e., selected model are not robust with respect to small perturbation of the sample)[3,4]. Post-selection inference issues also arise, such as underestimated standard errors and invalid confidence intervals, especially when variable selection and statistical inference are conducted on the same data, potentially compromising scientific conclusions[5].

Furthermore, variable selection is closely relate to the concept of "events per variable" (EPV). Particularly, in logistic regression, EPV refers to the ratio between the number of events and the number of predictors in the model. A popular rule of thumb recommends that EPV should be ≥ 10 to ensure stable selection[5–7].

Variable selection can be addressed through several approaches. Stepwise procedures are widely used for their simplicity, but are prone to overfitting, inflated Type I error, and instability[8]. Regularization techniques (e.g. LASSO[9],Elastic Net[10]) shrink coefficients toward zero to enable automatic selection, while. machine learning approaches yield variable importance measure with less interpretable selection criteria (e.g. Random Forests[11]). Heuristic search algorithms have also gained popularity for framing variable selection as an optimization problem[12,13].

Simulation studies are a fundamental tool in methodological research and provide a robust framework for evaluating the behavior and the properties of statistical methodologies. By generating data under controlled conditions and evaluating outcomes over repeated iterations, researchers can systematically compare the performance of variable selection algorithms assessing selection accuracy, model parsimony, and predictive performance under realistic constraints[14].

While many previous studies have focused on high-dimensional data or emphasized predictive performance alone, less attention has been given to a comprehensive, objective comparison of variable selection methods in low-dimensional clinical-like settings, where sample sizes are modest and variable interpretability is crucial. Yet, this is common in medical research and applied sciences.

To address this gap, we conducted a simulation study to compare five variable selection models in a binary classification framework, across a range of simulated scenarios encompassing varying key factors such as sample size, number of predictors, effect size and the strength of the correlation among variables. Our goal was to provide an assessment of these methods in terms of their selection accuracy, model sparsity, stability, and predictive performance, producing practical insights for researchers working with typical clinical data.

2. Simulation design–improvement through the ADEMP structure

To ensure a systematic and transparent simulation design, we followed the ADEMP structure proposed by Morris et al[14]. This involves specifying the Aim (A), Data-generating process (D), Estimands and other targets (E), analytical Methods implemented (M), and the criteria used to evaluate the Performance (P) of models. A summary of the design is provided in Table 1.

Table 1
Summary of the simulation design following the ADEMP structure.

Aim	**Comparison of variables selection methods in a classification framework and evaluation of their performances across different simulated scenario**
Data-generating mechanism	• **n =** 200, 1000 • **m** = 20, 100 • **p** = 3 for m = 20, 10 for m = 100 • Block correlation structure: informative predictors **ρ** = 0.2, 0.5, 0.8; non-informative predictors **ρ** = 0.4; between informative and non-informative predictors **ρ** = 0.2, 0.8 • **s** (effect size): high 0.40–1.50, low 0.15–0.45 (depends on n, m) • $Y \sim Bernoulli\left(expit\left(X\beta\right)\right)$
Estimands and other targets	• Number of variables selected • Number of true predictors selected • Predictive performances
Methods	• Bidirectional stepwise logistic regression • LASSO logistic regression • Elastic Net logistic regression • Random Forest classifier with OOB error based backward elimination • Genetic Algorithm with penalized fitness function
Performance measures	• Number of variable selected (N) • True Positive Rate (TPR) • False Discovery Rate (FDR) • Area Under the Curve (AUC)

2.1 Aims

Our study aims to objectively compare the performance, including predictive accuracy and the ability to retrieve the true model, of various variable selection methods across a range of simulation scenarios in the presence of numerous noise variables, under varying sample sizes and degrees of correlation among predictors.

2.2 Data Generating Mechanism

We generated several scenarios to reflect different levels of complexity and correlation among predictors trying to mimic real life clinical data structures. We considered simulation scenarios defined by the sample size (**n**) and the number of continuous covariates (**m**). Informative predictors were chosen as the first **p** continuous variables. The covariates matrices $\varvec{X} \in \mathbb{R}^{n \times m}$ were drawn from a multivariate normal distribution with mean $\mu = 0$ and a correlation following the block structure with one for informative covariates and one for non-informative covariates (Figure S1). Different values of correlation were chosen to simulate low, moderate and high correlation between informative variables. The correlation between non-informative predictors was held fixed whereas the correlation between informative and non-informative variables was set either low or high. A small amount of random noise was introduced in the block correlation matrix to better reflect real-world variability.[15,16].

The effect size of the regression coefficients *β* corresponding to the non-zero variables, were drawn from a normal distribution with mean zero and scaled by a prespecified parameter (**s**), indicating the signal strength, allowing us to control the informativeness of non-zero predictors.

Finally, the binary outcome variable $\varvec{Y}$ was generated from a Bernoulli distribution, where success probabilities were derived by applying the expit function to a logistic linear predictor without intercept. As a result, the linear predictor is symmetrically distributed around zero, so that success probabilities are centered around 0.5, yielding an approximately balanced binary outcome by construction.

2.3 Estimands and other targets

We evaluated several estimands capturing both variable selection accuracy and predictive performance. Selection accuracy was assessed through the inclusion of informative predictors and the exclusion of non-informative ones, as well as through exact recovery of the true model (i.e., selection of only the informative variables, with no false inclusions). We also examined the total number of variables selected, as an indicator of model parsimony versus complexity, and the predictive performance of the resulting models. Together, these estimands provide a comprehensive picture of each method's strengths and limitations in variable selection and classification tasks.

2.4 Methods

2.4.1 Stepwise logistic regression

This classical approach is a test-based variable selection method in which predictors are sequentially added or removed from the model according to statistical criteria such as p-values, adjusted R^2, Akaike Information Criterion (AIC), or Bayesian Information Criterion (BIC). Despite its popularity, stepwise regression has several well-documented limitations. It can lead to overfitting, produce biased coefficient estimates, and inflate the Type I error rate due to repeated hypothesis testing [17]. Moreover, because it searches the discrete model space greedily, it is not guaranteed to converge to the globally optimal subset and often settles instead on a locally optimal one. [18].

2.4.2 Penalized logistic regression models

Penalized regression methods improve performance and interpretability by penalizing the magnitude of the coefficients, shrinking estimates toward zero. We employed LASSO (Least Absolute Shrinkage and Selection Operator) and Elastic Net regularization within a logistic regression framework. LASSO imposes L_1-norm penalty on the regression coefficients, with the estimator $\widehat{\beta}$ obtained by minimizing the penalized loss

$$min \sum_{i=1}^{n} \left(y_i - x_i^T \beta\right)^2 + \lambda \sum_{j=1}^{p} \left|\beta_j\right|$$

where λ controls the amount of shrinkage. For sufficient large λ, the L_1 penalty forces some coefficient exactly to zero, yielding sparse solutions that exclude uninformative predictors. However, the resulting estimates are biased even for large coefficients, and LASSO tends to select only one variable from groups of correlated predictors[10,19]. To overcome these limitations, Zou and Hastie[10] proposed the Elastic net method, which combine L_1 and L_2 penalties to balance sparsity and stability. Its estimator solves

$$min \sum_{i=1}^{n} \left(y_i - x_i^T \beta\right)^2 + \lambda \left[(1-\alpha) \sum_{j=1}^{p} \left|\beta_j\right| + \alpha \sum_{j=1}^{p} \beta_j^2\right]$$

where λ and $\alpha \in [0,1]$ control the overall penalty and the mixture of norms. The combined penalty inherits sparsity from the L_1 term, while the L_2 term encourages correlated predictors to be selected together.

2.4.3 Random Forest classifier with OOB error based backward elimination

We adopted a variable selection procedure based on Random Forest (RF) presented by Díaz-Uriarte and de Andrés[20,21]. The main goal is to identify a minimal subset of variables that ensures strong predictive performance while avoiding redundancy. The selection process is iterative. At each step, a Random Forest model is fitted, variable importance is assessed via the Gini importance measure and variables with the lowest importance are progressively eliminated. For each fitted forest, the Out-of-Bag Error (OOBE) is recorded using the Out-of-Bag (OOB) samples (i.e. observations that were not included in the bootstrap sample used to build each tree). After iterating through several rounds of variable elimination, the final model is selected as the smallest subset of variables whose OOBE falls within a user-defined threshold *u* of the minimum OOBE achieved across all models.

2.4.4 Genetic Algorithm

Genetic Algorithms (GAs) are heuristic search methods inspired by the principles of biological evolution and natural selection[12,13,22,23]. A random initial population of individuals, called chromosomes, is generated, each representing a candidate model whose genes encode the inclusion or exclusion of

individual features. Across generations, the population evolves through processes analogous to biological reproduction to optimize a defined fitness function. Since variable selection is inherently an optimization problem, it can be addressed using GAs.

We initialized a population matrix of candidate models with dimensions $c \times m$, where c is the population size and m is the total number of variables. Each chromosome was evaluated by fitting a logistic regression model and computing the AUC using 5-fold stratified cross-validation. To balance performance and parsimony, we adopted the penalized fitness function proposed by Johnson et al.[23]:

$$fitness = AUC + d - \frac{d - l}{m},$$

where *l* denotes the number of active (selected) variables, and *d* is the ROC fitness offset, computed as the difference between the largest (AUC_l) and smallest (AUC_s) models considered:

$$d = AUC_l - AUC_s.$$

Parent chromosomes were selected via tournament selection[24] and combined through single-point crossover, with mutation applied at low probability (bit flip). Elitism preserved the best individual at each generation, and part of the population (20%) was periodically replaced with random chromosomes to prevent stagnation. The GA ran for up to 100 generations or until convergence. Parameters and workflow of the GA are shown in Fig. 1.

2.4.5 Hyperparameter optimization

Hyperparameters of the considered methods, except for the stepwise logistic regression, were optimized by applying random search with 60 iterations. Following Bergstra and Bengio[25], this number of iterations provides a 95% probability of sampling at least one configuration within the top 5% of the search space. The list of optimized parameter is reported in Table S1.

2.5 Performance measures

To evaluate and compare the chosen methods, we considered performance metrics that capture both selection accuracy and classification effectiveness namely the true positive rate (TPR), the proportion of informative predictors correctly selected, and the false discovery rate (FDR), the number of selected variables that are not informative. Whereas, predictive performance were evaluated using a test dataset (70/30 split) by computing the AUC. For each scenario we calculated the median value, along with the interquartile range (*IQR*), of the metrics considered over 1000 Monte Carlo simulations.

Computations

Due to the high computational cost and the independence of individual simulation runs, we adopted a MapReduce-inspired parallelization strategy. MapReduce is a programming model for distributed and scalable data processing, based on two core functions: map and reduce. The map function processes

the input data, produces intermediate jobs which are executed using several available processing units, while the reduce function aggregates all the produced results[26,27]. In our simulation framework, once the simulations are performed, the reduce phase aggregates the results by scenario.

The simulations were implemented in Python (3.12.7) using the Multiprocess[28] and Joblib[29] libraries to enable parallel execution on multiprocessor/multicore hardware.

3. Results

We hypothesized 48 different simulation scenarios by varying the aforementioned key parameters to evaluate the performance of variables selection models, over 1000 Monte-Carlo simulations (thus creating 240'000 single independent run), in a binary classification framework.

Overall performances, calculated on all scenarios (Table 2), showed that all methods achieved acceptable and comparable predictive performance with the AUC ranging from 0.766 [0.742–0.801] for the Random forest to 0.821 [0.786–0.858] for the Elastic net. Looking at the True Positive Rate (TPR) values, the Stepwise was the worst model, including half of the relevant predictors (TPR = 0.50 [0.37–0.63]), while the Elastic net correctly selected all of them (TPR = 1.00 [1.00–1.00]).

Table 2
Overall performance of variable selection methods over the 48 simulated scenarios. Values are reported ad median and IQR.

Method	Performance measures	
	AUC	TPR
Stepwise	0.79 [0.772–0.82]	0.5 [0.367–0.634]
LASSO	0.814 [0.78–0.851]	0.9 [0.667-1.0]
ElasticNet	0.824 [0.786–0.86]	1.0 [1.0–1.0]
RF	0.774 [0.746–0.816]	0.667 [0.6–0.8]
GA	0.786 [0.77–0.81]	0.6 [0.45–0.684]

The results in terms of variable selection accuracy and the predictive performance in each scenario are reported in Fig. 2–5.

The first 24 scenarios where built under n = 200, while scenarios 25–48 used n = 1,000. Within each sample-size block, two levels of m were considered: m = 20 with p = 3 (Scenarios 1–12 and 25–36) and m = 100 with p = 10 (Scenarios 13–24 and 37–48, respectively). Each block was further divided according to correlation and signal strength. For a detailed description of each simulation scenario see Table S2 in Supplementary. It should be noted that within each pair of adjacent scenarios (e.g., Scenarios

1 and 2, 7 and 8), all parameters remain constant, only the correlation between informative and non-informative variables varies.

Focusing on predictive performance (Fig. 2), AUC values were comparable across methods, with no model consistently outperforming the others.. Nonetheless, the Elastic Net achieved the highest AUC values overall, while the Random Forest exhibited the lowest. Predictive performance was higher in scenarios with m = 100 than with m = 20, with no clear pattern related to correlation. Higher AUC values were also observed when n = 1000 (Scenarios 25–48) and, within pairs of adjacent scenarios, when the correlation between informative and non-informative variables was 0.8.

Elastic net selected the largest number of variables, particularly in the *m* = 20 scenarios, whereas the Genetic Algorithm selected the most variables in 16 of the 24 m = 100 scenarios.. The fewest variables were selected by Stepwise logistic regression (m = 20) and Random Forest (m = 100) (Fig. 3).

Figure 4 shows that Elastic Net achieved high TPR values overall, successfully selecting all relevant predictors in nearly every simulated scenario. A result directly attributable to the large number of variables it selected. In contrast, both stepwise logistic regression and Random Forest struggled to include the non-zero predictors, selecting none of them in some scenarios. TPR values also generally increased in the *n* = 1000 scenarios.

Due to the large number of variables selected, Elastic Net and Genetic Algorithm showed high median FDR values. Stepwise logistic regression, despite selecting fewer variables than other methods, tended to select non-relevant predictors, particularly in low sample size scenarios. Random Forest selected the fewest non-relevant variables overall (Fig. 5).

Table S3 reports, for each performance measures, the median values and corresponding IQR achieved by the best-performing model in each scenario. In summary, Elastic Net exhibited the best AUC in 77% of scenarios and the best TPR in 50%. Stepwise logistic regression and Random Forest selected the fewest variables in 42% and 58% of scenarios, respectively, while Random Forest also showed the lowest FDR in 63% of scenarios (Table 3).

Table 3
Number and percentage of scenarios in which each variable selection method achieved the best performance, evaluated across four metrics (N, AUC, TPR, FDR).

Method	Performance measures			
	N	AUC	TPR	FDR
Stepwise	26 (54%)	4 (8%)	10 (21%)	17 (35%)
LASSO	-	2 (4%)	4 (8%)	1 (2%)
Elastic net	-	42 (87%)	28 (58%)	2 (4%)
RF	22 (46%)	-	6 (13%)	27 (56%)
GA	-	-	-	1 (2%)

4. Discussion

We conducted a comprehensive simulation study to compare five variable selection methods, some of them widely used in logistic regression modelling for clinical research. Across forty-eight scenarios, we examined how these methods performed under different data structures, varying degrees of association among true predictors and between true and false predictors, and predictors with both large and small effect sizes.

Our findings revealed consistent performance patterns across methods. Elastic Net emerged as the best-performing model in terms of AUC in nearly every scenario, also achieving high TPR values. However, it lacked parsimony, consistently selecting a large number of variables, a behavior attributable to its tendency to simultaneously select groups of correlated variables. LASSO achieved competitive performance in both prediction and parsimony. The Genetic Algorithm identified competitive subsets but tended to select larger models with high FDR, particularly under weak signal conditions, and required substantial computational effort. Random Forest produced sparse models while achieving the lowest, though still reasonable, predictive accuracy. Finally, Stepwise selection was suited for low-dimensional settings but frequently failed to recover the full set of informative variables, often including spurious ones, reinforcing longstanding criticisms .

Previous studies have evaluated variable selection methods in clinical settings[30–32]. Hanke and colleagues compared Best Subset Selection with forward stepwise selection, LASSO, and Elastic Net in a linear regression framework, reporting, consistently with our findings, superior performance by Elastic Net in both low- and high-dimensional contexts, particularly when predictors were correlated[31].

Sánchez-Pinto et al.[30] compared eight regression- and tree-based selection methods using two real-world clinical datasets of different sizes. In line with our results, tree-based approaches yielded sparser models and performed well in larger datasets, while penalized regression methods performed better in

smaller datasets with fewer events per variable (EPV). Similar findings for Random Forest were echoed by Van der Ploeg and Steyerberg[33]. In another study, Van der Ploeg et al.[32] found that classic logistic regression performed better in dataset with an EPV equals to 20–50, whereas Random Forest and other more modern approaches required an EPV > 200 to achieve stability. Unfortunately, most of our simulated scenarios featured EPV < 10, a known limitation for variable selection reliability[5]. Nevertheless, we ensured class balance in all simulations, which helps mitigate instability, as unbalanced outcome classes can bias variable selection, particularly for traditional methods[34].

Genetic Algorithms are designed for complex optimization problems and are well-suited for exploring vast model spaces when conventional methods struggle[12]. Several studies have reported promising results for GAs in feature selection tasks[23,35]. In our study, however, the GA often selected large models and underperformed in high-dimensional scenarios possibly reflecting insufficient exploration of the search space.

Our study has several limitations. First, our findings were derived entirely from synthetic datasets generated under idealized, fully controlled conditions. While simulations are valuable for method benchmarking, real-world data are typically more complex, featuring non-normal predictors, categorical variables, measurement error, and missing values. Evaluating these methods on real or semi-synthetic datasets (i.e. using correlation structure derived from actual data) would improve the generalizability of our results.

Second, we did not examine the properties of the estimated regression coefficients. Since valid statistical inference is a central concern in clinical research future work should also address this aspect.

Third, despite the wide range of scenarios, future research should expand to include additional methods such as stability selection, recursive feature elimination and boosting algorithms, as well as alternative correlation structures, varying EPV and class imbalance, to provide a more comprehensive understanding.

5. Conclusions

In summary, this simulation study compared the variable selection performance of five statistical methods within a binary classification framework, offering practical guidance for applied researchers and clinicians. Across 48 scenarios, all methods achieved comparable overall performance, yet no single method performed optimally in every scenario, and each came with distinct limitations. Future research should extend this comparison to real or semi-synthetic datasets and additional scenarios to support more generalizable recommendations.

Declarations

Authors Information

- Samuele Minari: s.minari@campus.unimib.it (**Corresponding author**)
- Antonella Zambon: antonella.zambon@unimib.it
- Dario Pescini: dario.pescini@unimib.it
- Mirko Cesarini: mirko.cesarini@unimib.it
- Davide Soranna: d.soranna@auxologico.it

Author Contributions

Conceptualization: Samuele Minari, Antonella Zambon.

Formal analysis: Samuele Minari.

Methodology: Samuele Minari, Antonella Zambon, Dario Pescini, Mirko Cesarini.

Software: Samuele Minari.

Supervision: Antonella Zambon.

Writing - original draft: Samuele Minari.

Writing - Review and editing: Antonella Zambon, Dario Pescini, Mirko Cesarini, Davide Soranna.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work the author(s) used Claude (Anthropic) for code development assistance and language editing. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the published article.

Founding declaration

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Data Availability

The datasets analysed in this study were generated by simulation. The code used to generate the simulated data and to perform all analyses is available in the GitHub repository at https://github.com/Siranim/Variable-selction-simulation-code.

References

1. Tolles J, Meurer WJ. Logistic Regression: Relating Patient Characteristics to Outcomes. JAMA. 2016;316(5):533–4. 10.1001/jama.2016.7653.

2. Kiezun A, Lee ITA, Shomron N. Evaluation of optimization techniques for variable selection in logistic regression applied to diagnosis of myocardial infarction. Bioinformation. 2009;3(7):311–3. 10.6026/97320630003311.
3. Ullmann T, Heinze G, Hafermann L, Schilhart-Wallisch C, Dunkler D, for TG2 of the STRATOS initiative. Evaluating variable selection methods for multivariable regression models: A simulation study protocol. Tian S. ed PLOS ONE. 2024;19(8):e0308543. 10.1371/journal.pone.0308543.
4. Sauerbrei W, Buchholz A, Boulesteix AL, Binder H. On stability issues in deriving multivariable regression models. Biom J. 2015;57(4):531–55. 10.1002/bimj.201300222.
5. Heinze G, Wallisch C, Dunkler D. Variable selection – A review and recommendations for the practicing statistician. Biom J. 2018;60(3):431–49. 10.1002/bimj.201700067.
6. Harrell FE, Lee KL, Matchar DB, Reichert TA. Regression models for prognostic prediction: advantages, problems, and suggested solutions. Cancer Treat Rep. 1985;69(10):1071–7.
7. Courvoisier DS, Combescure C, Agoritsas T, Gayet-Ageron A, Perneger TV. Performance of logistic regression modeling: beyond the number of events per variable, the role of data structure. J Clin Epidemiol. 2011;64(9):993–1000. 10.1016/j.jclinepi.2010.11.012.
8. Austin PC, Tu JV. Automated variable selection methods for logistic regression produced unstable models for predicting acute myocardial infarction mortality. J Clin Epidemiol. 2004;57(11):1138–46. 10.1016/j.jclinepi.2004.04.003.
9. Tibshirani R. Regression Shrinkage and Selection via the Lasso. J R Stat Soc Ser B Methodol. 1996;58(1):267–88.
10. Zou H, Hastie T. Regularization and Variable Selection Via the Elastic Net. J R Stat Soc Ser B Stat Methodol. 2005;67(2):301–20. 10.1111/j.1467-9868.2005.00503.x.
11. Breiman L. Random Forests. Mach Learn. 2001;45(1):5–32. 10.1023/A:1010933404324.
12. Zhang Z, Trevino V, Hoseini SS, et al. Variable selection in Logistic regression model with genetic algorithm. Ann Transl Med. 2018;6(3):45. 10.21037/atm.2018.01.15.
13. Escalona-Vargas D, Murphy P, Lowery CL, Eswaran H. Genetic algorithms for dipole location of fetal magnetocardiography. Annu Int Conf IEEE Eng Med Biol Soc IEEE Eng Med Biol Soc Annu Int Conf. 2016;2016:904–7. 10.1109/EMBC.2016.7590847.
14. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. Stat Med. 2019;38(11):2074–102. 10.1002/sim.8086.
15. Hardin J, Garcia SR, Golan D. A method for generating realistic correlation matrices. Ann Appl Stat. 2013;7(3). 10.1214/13-AOAS638.
16. Bag S, Gupta K, Deb S. A review and recommendations on variable selection methods in regression models for binary data. arXiv Preprint posted online May. 2022;16:220106063. 10.48550/arXiv.2201.06063.
17. Harrell FE. Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis. Springer International Publishing; 2015. 10.1007/978-3-319-

19425-7.

18. Zou H. The Adaptive Lasso and Its Oracle Properties. J Am Stat Assoc. 2006;101(476):1418–29. 10.1198/016214506000000735.
19. Zhang S, Zhang L, Qiu K, Lu Y, Cai B. Variable Selection in Logistic Regression Model. Chin J Electron. 2015;24(4):813–7. 10.1049/cje.2015.10.025.
20. Diaz-Uriarte R, de Andres SA. Variable selection from random forests: application to gene expression data. *arXiv*. Preprint posted online June 22, 2005:arXiv:q-bio/0503025. 10.48550/arXiv.q-bio/0503025
21. Díaz-Uriarte R, Alvarez de Andrés S. Gene selection and classification of microarray data using random forest. BMC Bioinformatics. 2006;7(1):3. 10.1186/1471-2105-7-3.
22. Melanie M. An Introduction to Genetic Algorithms.
23. Johnson P, Vandewater L, Wilson W, et al. Genetic algorithm with logistic regression for prediction of progression to Alzheimer's disease. BMC Bioinformatics. 2014;15(16):S11. 10.1186/1471-2105-15-S16-S11.
24. Miller BL. Genetic Algorithms, Tournament Selection, and the Effects of Noise.
25. Bergstra J, Bengio Y. Random Search for Hyper-Parameter Optimization. J Mach Learn Res. 2012;13(10):281–305.
26. Dean J, Ghemawat S. MapReduce: simplified data processing on large clusters. Commun ACM. 2008;51(1):107–13. 10.1145/1327452.1327492.
27. Hashem IAT, Anuar NB, Gani A, Yaqoob I, Xia F, Khan SU. MapReduce: Review and open challenges. Scientometrics. 2016;109(1):389–422. 10.1007/s11192-016-1945-y.
28. multiprocessing — Process-based parallelism. Python documentation. Accessed March 23. 2026. https://docs.python.org/3/library/multiprocessing.html
29. developers T joblib. joblib. Published online Dec. 2025;15. 10.5281/zenodo.17936197.
30. Sanchez-Pinto LN, Venable LR, Fahrenbach J, Churpek MM. Comparison of variable selection methods for clinical predictive modeling. Int J Med Inf. 2018;116:10–7. 10.1016/j.ijmedinf.2018.05.006.
31. Hanke M, Dijkstra L, Foraita R, Didelez V. Variable selection in linear regression models: Choosing the best subset is not always the best choice. Biom J. 2024;66(1):2200209. 10.1002/bimj.202200209.
32. van der Ploeg T, Austin PC, Steyerberg EW. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. BMC Med Res Methodol. 2014;14(1):137. 10.1186/1471-2288-14-137.
33. van der Ploeg T, Steyerberg EW. Feature selection and validated predictive performance in the domain of Legionella pneumophila: a comparative study. BMC Res Notes. 2016;9:147. 10.1186/s13104-016-1945-2.

34. Kamalov F, Thabtah F, Leung HH. Feature Selection in Imbalanced Data. Ann Data Sci. 2023;10(6):1527–41. 10.1007/s40745-021-00366-5.
35. Gayou O, Das SK, Zhou SM, Marks LB, Parda DS, Miften M. A genetic algorithm for variable selection in logistic regression analysis of radiotherapy treatment outcomes. Med Phys. 2008;35(12):5426–33. 10.1118/1.3005974.

Figures

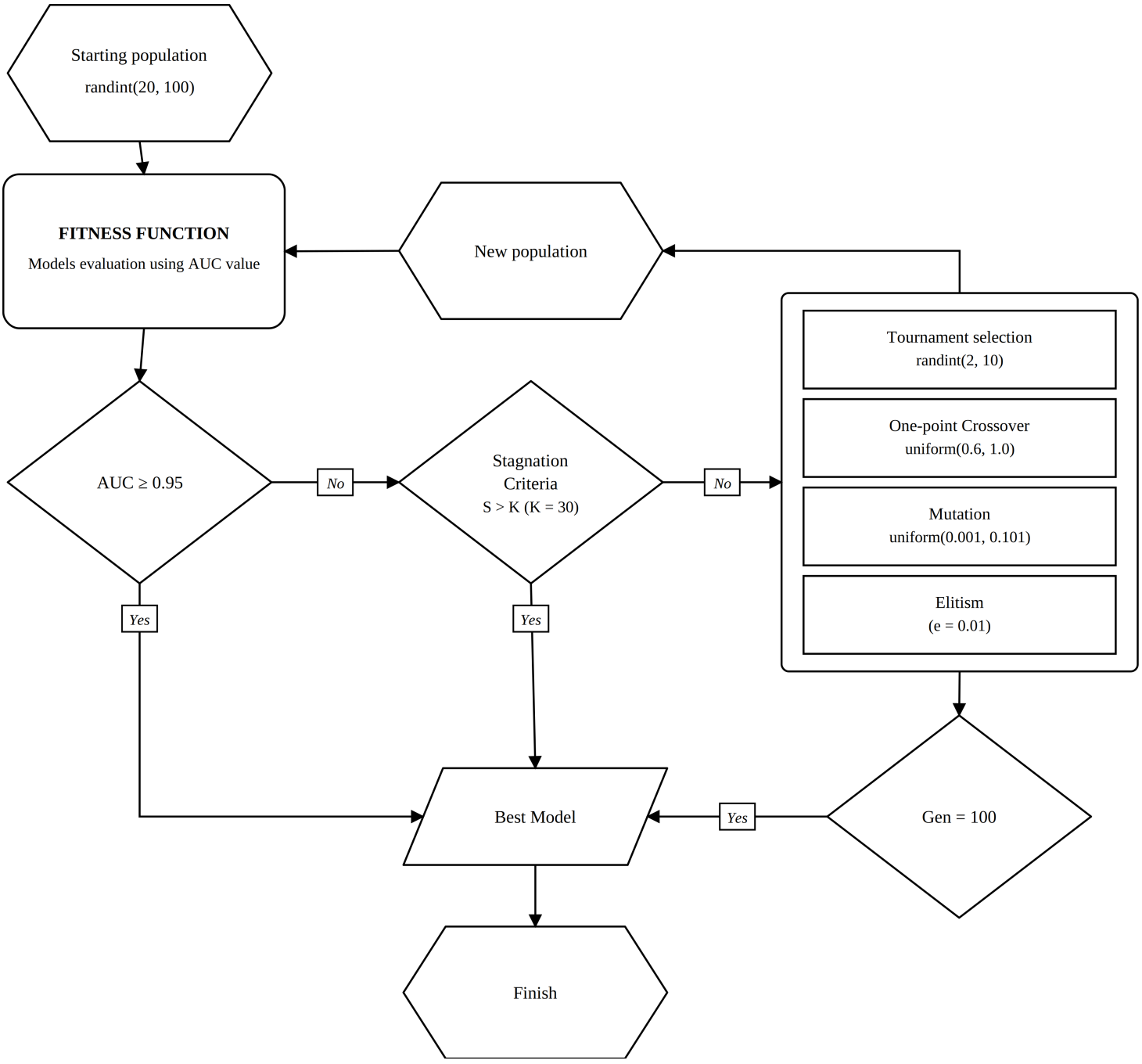


Figure 1

Diagram representing the Genetic Algorithm (GA) structure.

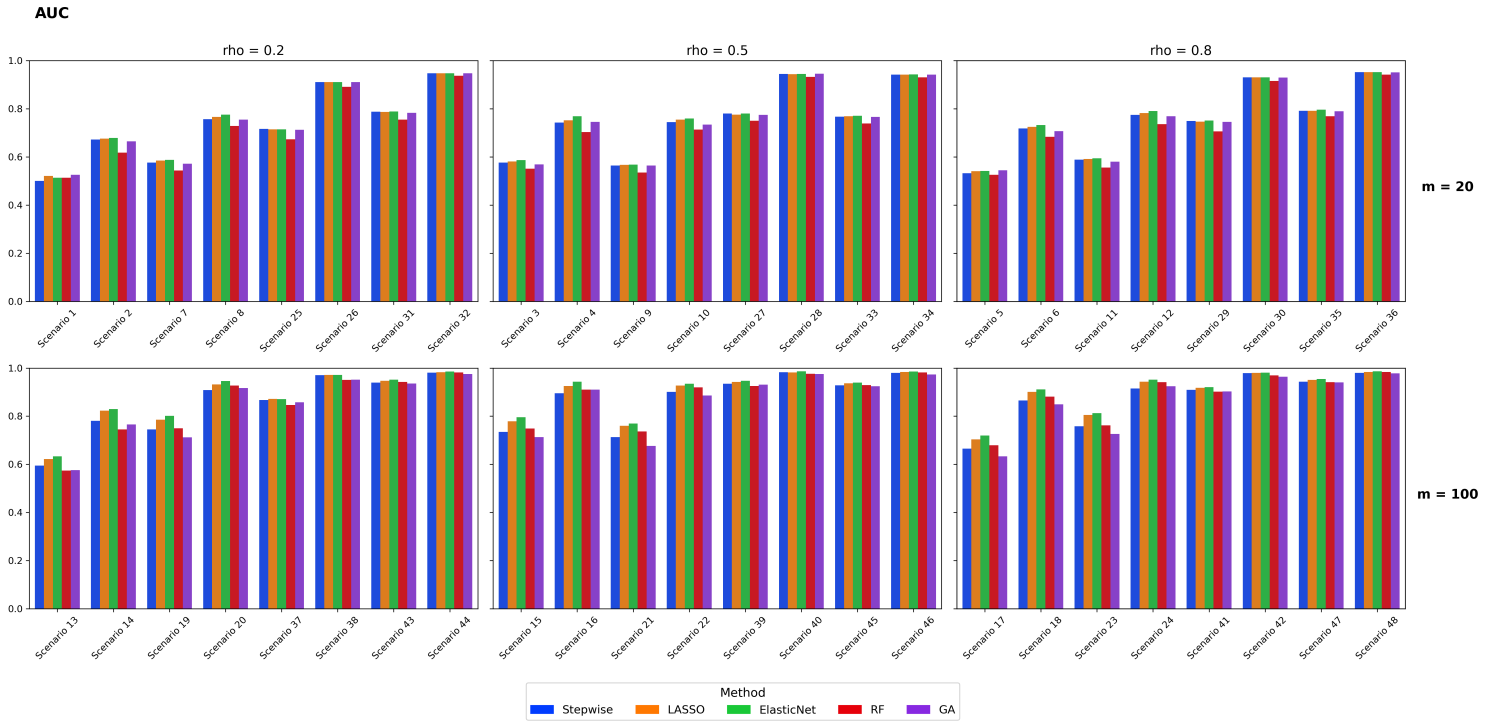


Figure 2

Median AUC values for each variable selection method (Stepwise, LASSO, Elastic Net, RF, GA) across all simulation scenarios.

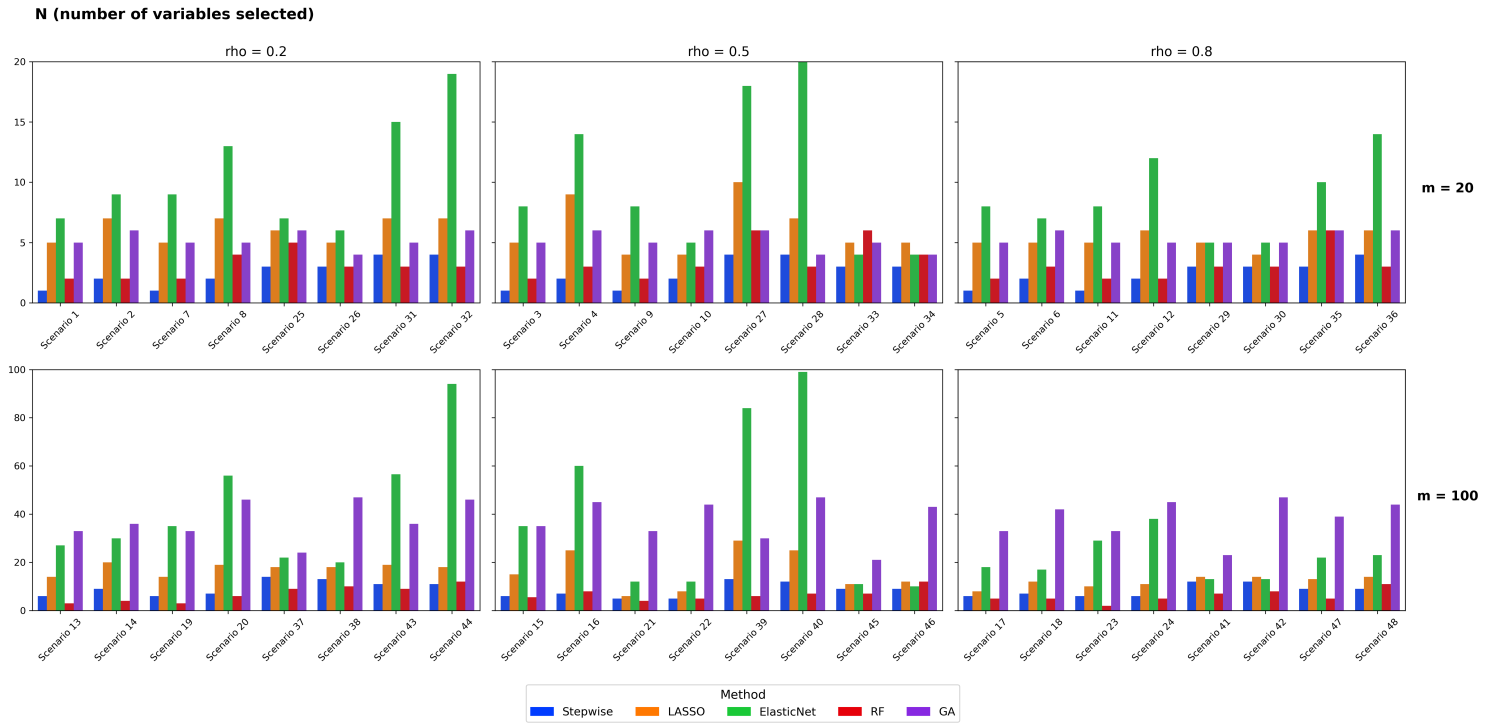


Figure 3

Median number of variables selected for each variable selection method (Stepwise, LASSO, Elastic Net, RF, GA) across all simulation scenarios.

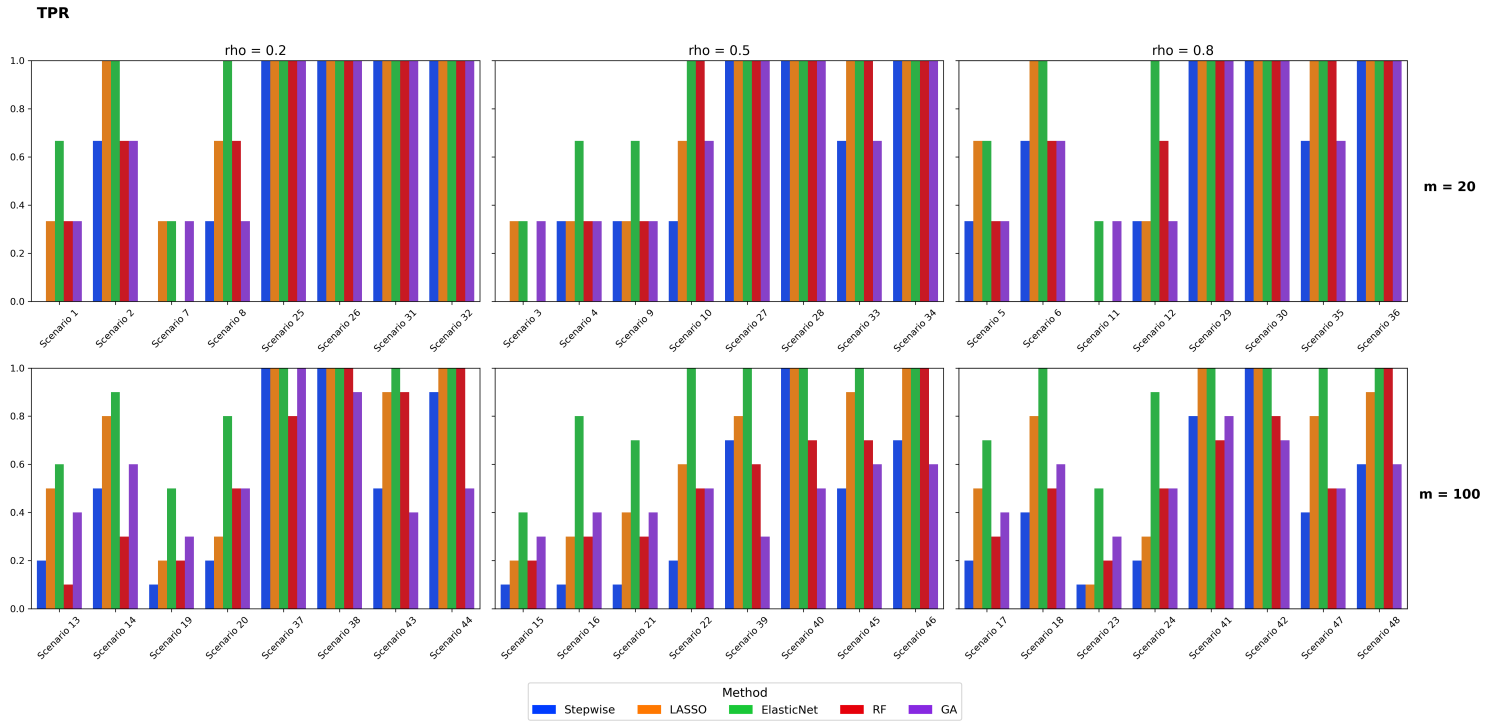


Figure 4

Median True Positive Rate (TPR) values for each variable selection method (Stepwise, LASSO, Elastic Net, RF, GA) across all simulation scenarios.

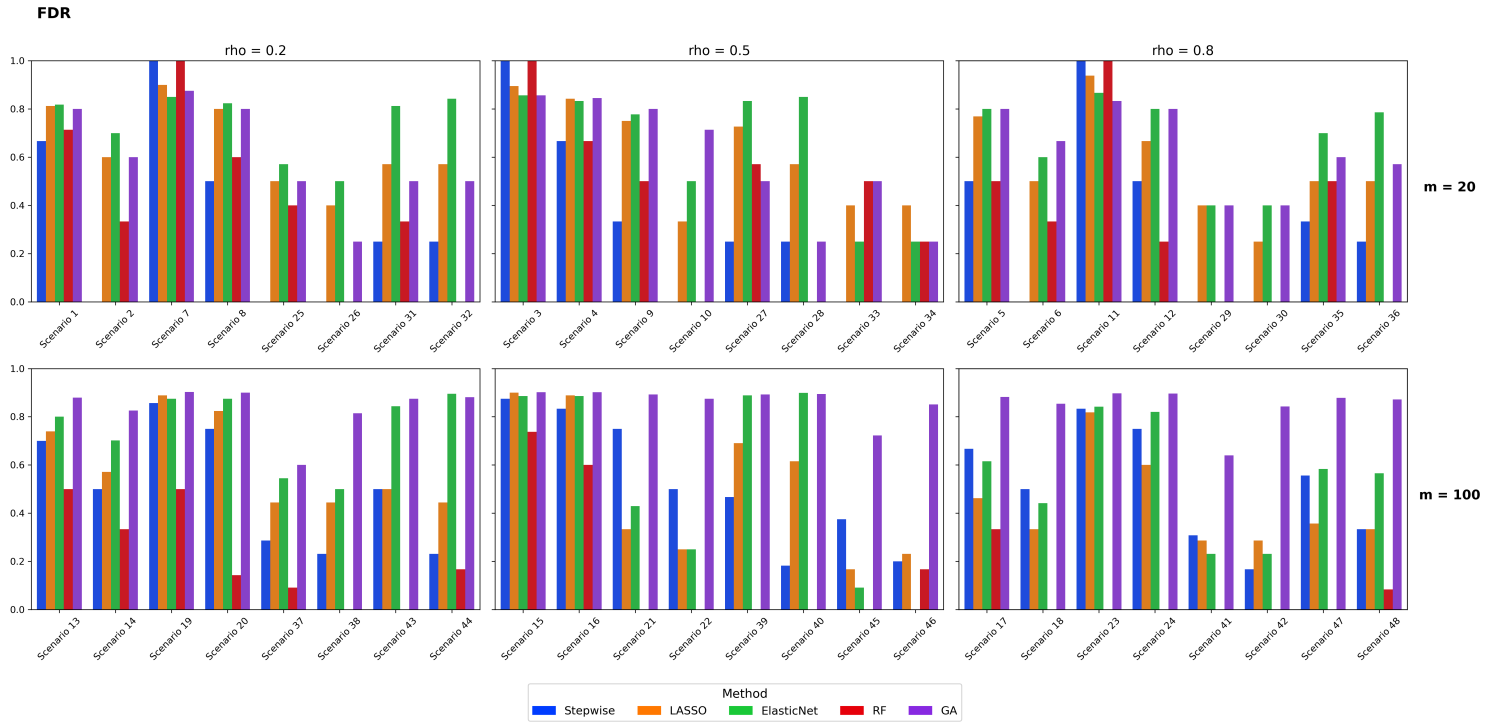


Figure 5

Median False Discovery Rate (FDR) values for each variable selection method (Stepwise, LASSO, Elastic Net, RF, GA) across all simulation scenarios.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementary.docx