Original Research Article

MDM Medical Decision Making

Medical Decision Making
2023, Vol. 43(1) 68–77

DOI: 10.1177/0272989X221125418
journals.sagepub.com/home/mdm
SAGE

Cost-Effectiveness and Value-of-Information Analysis Using Machine Learning–Based Metamodeling: A Case of Hepatitis C Treatment

John Austin McCandlish, Turgay Ayer, and Jagpreet Chhatwal

Background. Metamodels can address some of the limitations of complex simulation models by formulating a mathematical relationship between input parameters and simulation model outcomes. Our objective was to develop and compare the performance of a machine learning (ML)–based metamodel against a conventional metamodeling approach in replicating the findings of a complex simulation model. **Methods.** We constructed 3 ML-based metamodels using random forest, support vector regression, and artificial neural networks and a linear regression-based metamodel from a previously validated microsimulation model of the natural history hepatitis C virus (HCV) consisting of 40 input parameters. Outcomes of interest included societal costs and quality-adjusted life-years (QALYs), the incremental cost-effectiveness (ICER) of HCV treatment versus no treatment, cost-effectiveness analysis curve (CEAC), and expected value of perfect information (EVPI). We evaluated metamodel performance using root mean squared error (RMSE) and Pearson's R^2 on the normalized data. **Results.** The R^2 values for the linear regression metamodel for QALYs without treatment, QALYs with treatment, societal cost without treatment, societal cost with treatment, and ICER were 0.92, 0.98, 0.85, 0.92, and 0.60, respectively. The corresponding R^2 values for our ML-based metamodels were 0.96, 0.97, 0.90, 0.95, and 0.49 for support vector regression; 0.99, 0.83, 0.99, 0.99, and 0.82 for artificial neural network; and 0.99, 0.99, 0.99, 0.99, and 0.98 for random forest. Similar trends were observed for RMSE. The CEAC and EVPI curves produced by the random forest metamodel matched the results of the simulation output more closely than the linear regression metamodel. **Conclusions** ML-based metamodels generally outperformed traditional linear regression metamodels at replicating results from complex simulation models, with random forest metamodels performing best.

Highlights

- Decision-analytic models are frequently used by policy makers and other stakeholders to assess the impact of new medical technologies and interventions. However, complex models can impose limitations on conducting probabilistic sensitivity analysis and value-of-information analysis, and may not be suitable for developing online decision-support tools.
- Metamodels, which accurately formulate a mathematical relationship between input parameters and model outcomes, can replicate complex simulation models and address the above limitation.
- The machine learning–based random forest model can outperform linear regression in replicating the findings of a complex simulation model. Such a metamodel can be used for conducting cost-effectiveness and value-of-information analyses or developing online decision support tools.

Corresponding Author
Jagpreet Chhatwal, MGH Institute for Technology Assessment 101 Merrimac Street, Suite 1010, Boston, MA 01887, USA; (jagchhatwal@mgh.harvard.edu).

Keywords
cost-effectiveness, machine learning, metamodels, microsimulation, sensitivity analysis, value of information analysis

Date received: February 02, 2022; accepted: August 22, 2022

Introduction

Policy makers and other stakeholders frequently use decision-analytic models to assess the impact of new medical technologies and interventions.[1] Recently, with the availability of more data and computational speed, the complexity of decision-analytic models has increased.[2] Such computationally intensive simulation models are useful for accurately capturing important real-world nuances in disease modeling.

However, the increased complexity of simulation models can impose limits on conducting probabilistic sensitivity analysis (PSA) and value-of-information analysis (VOI),[1] which are recommended by various cost-effectiveness guidelines to identify future research priorities.[3] For instance, the Agency for Healthcare Research and Quality published a report on VOI that included a systematic review of 60 studies conducting model-based VOI analysis,[4] and out of the 60 studies, only 3 used microsimulation modeling.[5–7] Furthermore, complex models may not be suitable for developing online decision-support tools needed by different stakeholders, because complex models may not generate outcomes in near real time.

To address these challenges, researchers use metamodels, which formulate a mathematical relationship between input parameters and simulation model outcomes. Metamodels are sometimes called models of models, and a well-calibrated metamodel may replace the original simulation model when timely results are needed. In addition to efficiently conducting extensive sensitivity analyses and VOI, metamodels can also be used for developing online decision-support tools for real-time use.[8,9] This makes metamodels a vital tool for researchers seeking fast and reliable information necessary for decision making.

Traditional metamodeling techniques, such as linear regression metamodels (LRMs) and Gaussian process metamodels (GPMs), have limitations.[9] LRMs are often ineffective at modeling nonlinear relationships between outcomes and parameters, which is typically the case when dealing with complex simulation models. Adjusting LRMs to capture such nonlinear relationships quickly becomes computationally expensive and increases the potential for structurally biased estimates.[10] GPMs, while quite effective at capturing nonlinear trends, are computationally expensive and require advanced statistical expertise for tuning to accurately capture the relationship between input parameters and outcomes.[11] Of note, only a limited number of existing metamodeling techniques have been used for cost-effectiveness analysis and VOI.[12]

Machine learning (ML)–based metamodels can offer some advantages over the commonly used LRMs and GPMs. ML models have superior predictive accuracy over traditional statistical methods when the functional form being modeled exhibits high degrees of nonlinearity or is noncontinuous. This is especially applicable to cost-effectiveness analyses, in which the outputs (e.g., incremental cost-effectiveness ratio) are nonlinear and highly sensitive to model inputs. While there are numerous examples of ML-based metamodels in other fields such as information systems,[13–15] there are few examples of ML-based metamodels in medical decision making.[9,16,17] The application of ML methods in health economics and outcomes research (HEOR) remains in the early stages.[9]

The objective of this study is to compare the performance of 3 ML-based metamodels—random forest (RF), support vector regression (SVR), and artificial neural network (ANN) metamodels—with a commonly used LRM in replicating a complex microsimulation model, the Markov-based Analysis of Treatment for Chronic Hepatitis C (MATCH),[18–20] designed for analyzing the cost-effectiveness of direct-acting antivirals, a new treatment for hepatitis C. We further compare the performance of these different metamodels for conducting VOI analysis and replicating cost-effectiveness analysis. This case study aims to demonstrate the ability of ML-based metamodels to replicate a highly complex simulation

Georgia Institute of Technology, Atlanta, Georgia (JAM, TA); Massachusetts General Hospital Institute for Technology Assessment, Boston, Massachusetts (JC); Harvard Medical School, Boston, Massachusetts (JC). The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article. The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: National Science Foundation Award numbers 1722614 and 1722665. The funding agreement ensured the authors' independence in designing the study, interpreting the data, writing, and publishing the report.

model designed for HEOR and to assess their performance relative to traditional statistical methods in terms of accuracy, precision, and computation time.

Methods

Rationale for the Choice of Metamodels

Linear regression metamodels (LRMs) are intuitive and have strong theoretical guarantees if formulated correctly.[8] The inclusion of coefficients, confidence intervals, and *P* values for each input parameter allows for approximations of parameter contribution and uncertainty.[8] However, LRMs have considerable limitations. First and foremost, the strict parametric assumptions of LRMs make them highly sensitive to model specification, which must be stipulated by the researcher. An improper formulation is shown to produce highly biased predictions and coefficient estimates.[10] Furthermore, the estimation of nonlinearities and interactions between input parameters involves a significant expansion of the parameter space, increasing computation time. A simple full second-order model with $p-1$ input parameters requires estimation of $2p + \binom{p}{2}$ coefficients; thus, a full second-order LRM for a simulation model with 40 input parameters would require the estimation of 902 coefficients.[21] Yet, past studies have found that inclusion of interaction and higher-order effects are often necessary when constructing LRMs for complex simulation models, especially for conducting cost-effectiveness analysis.[8] In addition, the inclusion of higher-order terms also comes at the cost of interpretability, as estimating individual variable contributions requires consideration of all nonlinear terms and interactions.

GPMs are Bayesian nonparametric metamodels that exploit the spatial distance between model input and output values. GPMs are better equipped to estimate nonlinear relations than LRMs, and like LRMs, they have strong theoretical guarantees if properly formulated. The parametric assumptions of GPMs also ensure that the prediction function is smooth. However, this greater flexibility comes at a significant expense of computational efficiency and interpretability. Most current GP software is ill-equipped to handle models with more than 30 parameters and necessitates a high level of statistical expertise to ensure proper smoothing.[9,11] Further, training a GPM requires a substantial amount of computational time (often days or weeks to train on a data set of 10,000 observations).[11] Crucially, the complexity and parametric restrictions of GPMs can make VOI and threshold analysis difficult and time-consuming.[9] Given the number of input parameters in the MATCH model and the size of our PSA sample, we did not include a GPM for comparison in this study.

ML-based algorithms have received widespread attention in many other fields but have remained relatively unexamined in the metamodeling literature, especially within the context of medical decision making and HEOR.[9] There are many ready-to-implement ML algorithms widely available on free software packages such as R and Python. RF, SVR, and ANN models have all demonstrated high predictive accuracy in several applications when the functional form being modeled exhibits high degrees of nonlinearity and cannot be accurately modeled using traditional statistical methods.[21] Perhaps the biggest drawback of these models is their lack of interpretability: ANN models, for example, are pure black boxes, so they provide little insight into the underlying structure of the data. Nevertheless, many of these algorithms, such as RF, do maintain a degree of interpretability in addition to potentially valuable characteristics, such as measures of which input parameters drive outcomes or the ability to cluster like observations based on both input and output characteristics.

Hepatitis C Simulation Model

We use a previously validated microsimulation model, the MATCH, a comprehensive model that simulates the progression of chronic hepatitis C and estimates the cost-effectiveness of direct-acting antiviral treatment.[18–20] We developed different metamodels using the techniques described above (LR, RF, SVR, and ANN) designed to replicate the outcomes of MATCH.

In the MATCH model, patients progress through different stages of chronic hepatitis C (defined by fibrosis stages: F0 through F4), decompensated cirrhosis of the liver, hepatocellular carcinoma (liver cancer), and need for a liver transplant, with death being an absorbing state (Figure 1). Patients who receive hepatitis C treatment could achieve sustained virologic response, a surrogate for a cure. The model accounts for costs associated with the management of hepatitis C disease (e.g., liver transplant and liver cancer) and the cost of antiviral treatment, as well as quality-of-life adjustments for fibrosis, cirrhosis, and liver cancer. The MATCH model uses 40 input parameters and calculates the average total cost and average quality-adjusted life-years (QALYs) per patient for both treatment and no treatment with a lifetime horizon. The incremental cost-effectiveness ratio (ICER) of hepatitis C treatment is the key outcome of interest to researchers in HEOR. Uncertainty is captured via PSA, with each run varying 40 input parameters that

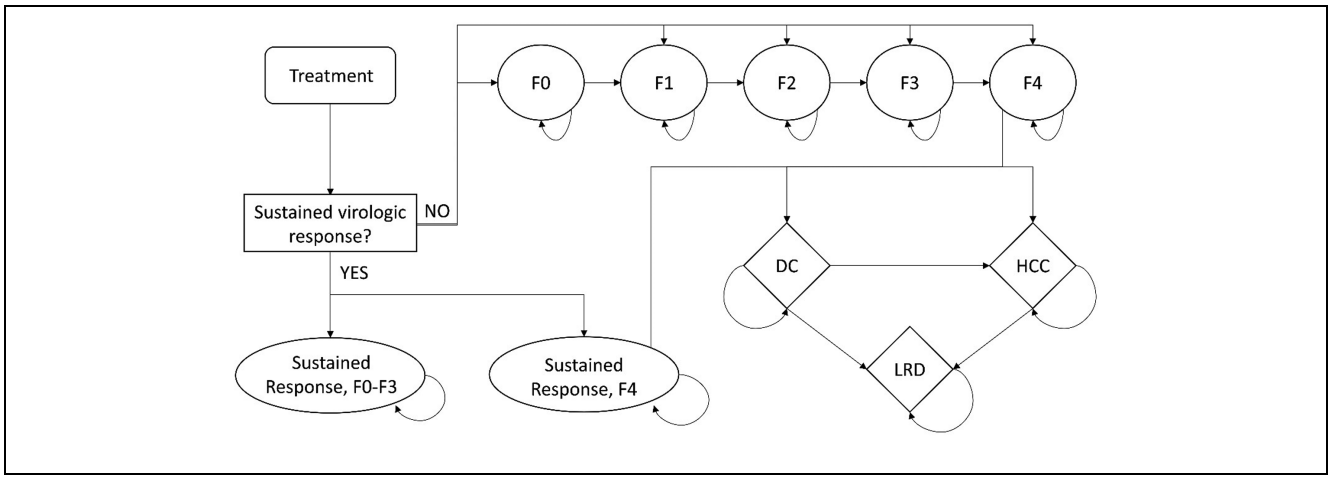

Figure 1 Schematic for the simulation model of hepatitis C natural history and treatment. The microsimulation model tracks the health and cost outcomes associated with progression of hepatitis C. At a given time, a patient occupies one of the health states. Arrows between states represent possible transitions based on annual probabilities. As time progresses, patients can transition to another state and acquire cost and health utilities associated with that state. The model stops when all patients transition to the death state. A patient could transition to a death state from any of the other states because of background mortality (these transitions are not shown for clarity). F0-F4 = METAVIR fibrosis stages; DC = decompensated cirrhosis; HCC = hepatocellular carcinoma; LRD = liver-related death; LT = liver transplantation. The DC and LT states were further divided into first-year and subsequent-year states to account for different mortality rates and costs; however, they are collapsed into 1 state for presentation purposes.

account for quality-of-life weights (10 parameters), costs (8 parameters), probabilities (18 parameters), and patient characteristics (4 parameters) according to predefined distributions (see Appendix Table 1 for a detailed description of each input parameter and its assigned PSA distribution). While the original MATCH model includes different types of hepatitis C treatment, we simulated only 2 strategies—antiviral treatment versus no treatment—when building the metamodel for simplicity.

Data-Generation Processes

The data set for metamodel construction is most often generated using either a PSA output or a deterministic design such as full-factorial SA design.[8] Our data-generation process involved using 1000 hypothetical patients for the inner loop and 1 million parameter draws for the outer loop. Our final PSA sample was produced from the original simulation model using C++ and took 1 wk to run on an Acer Aspire 5 with an Intel Core 7-7700 processor. For each PSA simulation, all but 5 parameters were drawn from gamma distributions (patient characteristics and treatment costs were drawn from uniform distributions). Our final data set contained 5 outcomes of interest: total QALYs with and without treatment, societal-level costs with and without treatment, and the ICER of hepatitis C treatment, as well as the realized values of all 40 input parameters for each simulation run. We chose to generate our data directly from PSA rather than a deterministic design because of its ability to generate a large number of observations, and because, as Jalal et al.[8] point out, the R^2 value can be misleading when deterministic designs are used because each parameter has only a few data points to capture variation. In addition, as PSA is usually a required step in the evaluation of HEOR models, it is a more convenient data-generation process than a full-factorial design.[8]

Construction of Metamodels

Three of the 40 input parameters of the MATCH model were categorical (patient sex, starting stage, and disease genotype). As such, we dichotomized these categorical variables by adding dummy variables for these parameters for all metamodels, bringing our total number of parameters to 48. Finally, we normalized all input parameters and outcomes and randomly divided the PSA sample into a training set of 100,000 simulations and a testing set of 900,000 simulations. While it is typical for the testing set to be smaller than the training set, a central benefit of constructing a metamodel is the reduced

computation time required to conduct complex analyses. For example, expected value of partial perfect information (EVPPI), which estimates the value of knowing a subset of input parameters, often requires hundreds of thousands of simulation scenarios. As such, we were interested in assessing the ability of metamodels to replicate a large PSA sample of unseen simulation scenarios. A larger testing set of unseen simulation scenarios also provides an indicative test of whether a metamodel has successfully re-created the data-generation process of the original simulation. We chose a training set size of 100,000 because it represents a PSA sample that can be produced in a reasonable time frame (approximately 17 h).

We constructed 2 LRM models for QALYs without treatment, QALYs with treatment, societal-level cost without treatment, societal-level cost with treatment, and ICER. We first fit a simple LRM that included only main effects for each parameter. To account for nonlinearity and potential interaction effects, we next fit a partial second-order model that contained the squared values of all features as well as all within-group interaction effects for probability, quality of life, cost, and patient characteristic parameters. Because the structure of the simulation model was such that QALYs and societal-level cost were each calculated omitting cost parameters and quality-of-life parameters, respectively, we did not include quality-of-life weights in our LRM for societal-level cost, nor did we include cost parameters in our LRM for QALYs. This brought the total number of features for our QALY second-order LRMs to 266, our cost second-order LRMs to 245, and our ICER second-order LRM to 310. We also fit LRMs with third-order terms for all parameters but found no improvement in R^2, and so the second-order LRMs were selected to be our baseline comparison metamodel. All LRMs were trained using R version 4.03 for Windows 10.[22]

An overview of ANN, RF, and SVR algorithms can be seen in Lazic.[23] All ML-based metamodels required tuning hyperparameters. The ANN metamodel required tuning of 4 hyperparameters: number of hidden layers (1, 2), nodes per layer (25, 50, 75, 100), activation function (identity, logistic, hyperbolic tan, rectified linear unit), and alpha value (0.0001, 0.001, 0.01). The RF metamodel required tuning 2 hyperparameters: number of trees (50, 100, 150, 200) and maximum number of features selected for splitting (10, 15, 20, 25, 30). The SVR metamodel required tuning 3 hyperparameters: kernel (linear, polynomial, radial, sigmoid), C (1, 10, 100), and epsilon (0.001, 0.01, 0.1). To reduce computation time, hyperparameters were tuned using a subsample of 10,000 PSA simulations with 5-fold cross-validation. The final set of hyperparameters for each metamodel was selected based on the average cross-validated R^2 values across outcomes using an exhaustive grid search across all 96, 25, and 36 possible hyperparameter combinations for ANN, RF, and SVR metamodels, respectively. A single set of hyperparameters was selected for all outcomes, and final models were trained on the full training set of 100,000 simulations. The RF metamodel was tuned and trained using the ranger package in R version 4.03 for Windows 10, while the ANN and SVR metamodels were tuned and trained using the scikit-learn in Python version 3.6 for Windows 10.[24,25]

Comparison of Outcomes

For each model, we used RMSE and R^2 on the normalized testing sample of 900,000 PSA simulations as the primary performance metrics for comparing the models. We also constructed a cost-effectiveness acceptability curve (CEAC) using both data from the original PSA testing sample and the predicted values from the LRM and the best-performing ML-based metamodel. Finally, we calculated the expected value of perfect information (EVPI), which quantifies how much a society is willing to pay to eliminate parametric uncertainty in the decision. Equations 2 and 3 show its calculation, as outlined by Briggs et al.[3,8–11]

$$EVPI = \mathbb{E}_\theta\{\max_x NB(x,\theta)\} - \max_x[\mathbb{E}_\theta\{NB(x,\theta)\}] \quad (2)$$

$$NB(x,\theta) = QALYs(x,\theta) * WTP - Cost(x,\theta), \quad \textit{for strategy } x \textit{ and input parameters } \theta \quad (3)$$

The EVPI was calculated for willingness-to-pay (WTP) thresholds from $1,000 to $200,000 using both the original PSA testing sample and the predicted values from the LRM and best-performing ML-based metamodel.

Financial support for this study was provided in part by a grant from the National Science Foundation. The funders played no role in the objectives, methods, or conclusions of this study.

Results

The amount of time required for hyperparameter tuning for each ML-based metamodel was 24 min for ANN, 3 h for RF, and 5 h for SVR. Training final metamodels for all 5 outcomes took 1, 3, 16, and 135 min for LR, ANN, RF, and SVR, respectively. All metamodels were able to generate predicted outcomes for our testing set of 900,000 PSA simulations in less than 1 min. The final

Table 1 Performance Metrics for Metamodels

Model	QALY: No Treatment	QALY: Treatment	Cost: No Treatment	Cost: Treatment	ICER[a]
LR					
RMSE	0.2753	0.1497	0.3877	0.2887	0.0069
R^2	0.924	0.978	0.850	0.917	0.601
RF					
RMSE	0.0793	0.0484	0.2387	0.1378	0.0048
R^2	0.994	0.998	0.943	0.981	0.810
ANN					
RMSE	0.0491	0.0468	0.0917	0.0729	0.0071
R^2	0.998	0.998	0.992	0.995	0.582
SVR					
RMSE	0.2968	0.1580	0.4500	0.3238	0.0078
R^2	0.912	0.975	0.797	0.895	0.489

ANN, artificial neural network; ICER, incremental cost-effectiveness ratio; LR, linear regression; QALY, quality-adjusted life year; RF, random forest; RMSE, root mean squared error; SVR, support vector regression. The RMSE and R^2 values reported for both the LRM and RF metamodel, with the true simulation output as the predictors. Separate metamodels were constructed for the 5 main outputs of the simulation model.

[a]Because the simulation model produced some extremely large ICER values, we report only the performance of our metamodels for ICER values between $0 and $300,000.

ANN metamodels had 2 hidden layers of 100 nodes each, a logistic activation function, and an alpha value of 0.001. The final RF metamodels had 200 trees and 30 candidate parameters for splitting. The final SVR metamodels had a linear kernel, a C value of 1.0, and an epsilon value of 0.1.

Table 1 shows the normalized RMSE and R^2 values for all metamodels. The SVR metamodels produced lower R^2 values (higher RMSEs) than the LRMs across all outcomes. The ANN metamodels produced higher R^2 values (lower RMSEs) than the LRMs for all outcomes apart from ICER, in which the R^2 values were comparable (R^2 of 0.582 compared with 0.601 for LRM). Our RF metamodels produced higher R^2 values (lower RMSEs) than the LRMs. Because it performed better than the LRM across outcomes and because it outperformed all other metamodels on ICER, we selected the RF metamodel as our primary comparator to the baseline LRM.

Of note, the R^2 values for ICER for all metamodels were uninterpretable when all 1 million observations were included for training, as both models produced negative R^2 values. Further analysis revealed that this could be attributed to rare outliers in ICER produced by the original simulation model. Both models performed poorly at predicting ICER values for observations in which the true ICER was greater than $300,000. This is likely because, in the original model, large ICER values were infrequently observed and quite variable. As ICER is estimated by dividing changes in cost by changes in QALYs, ICER values become quite large as the difference in QALYs between treatment and no treatment approaches zero. Researchers interested in analyzing the cost-effectiveness of interventions typically set WTP thresholds well below $300,000/QALY: the World Health Organization recommends an upper-bound WTP threshold of 3× gross domestic product per capita ($190,600), and the National Institute for Health and Care Excellence uses a threshold of £100,000/QALY ($131,000) for evaluation of highly specialized technologies.[26] Thus, the R^2 values for ICER in Table 1 were calculated using only those observations with true ICER values at or below $300,000.[25] We verified that both the RF metamodel and LRM predicted ICER values of greater than $300,000 for all observations, with true ICER values greater than $300,000. We decided this was a reasonable exclusion for 2 reasons: fewer than 1.3% of observations had ICER values of more than $300,000, and large prediction errors for ICER values of more than $300,000 would be unlikely to alter a cost-effectiveness determinations so long as predicted values were also beyond $300,000. In some cases, such as for rare diseases, a much higher threshold may be used, and we did not evaluate such cases as hepatitis C is not a rare disease.

Figure 2 shows a scatterplot of the predicted versus actual ICER values for the LRM and RF metamodel using the PSA testing sample. The RF metamodel predictions remained more centered and closely clustered around the line of fit than the LRM predictions across WTP thresholds from $1,000 to $200,000. LRM predictions demonstrated systematic bias, with ICER predictions systematically high for low ICER values and systematically low for high ICER values.

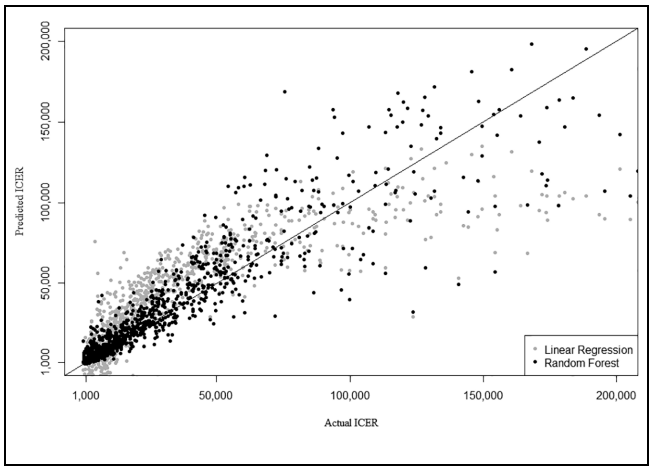

Figure 2 Actual versus predicted incremental cost-effectiveness ratio values for the linear regression metamodel and random forest metamodel.

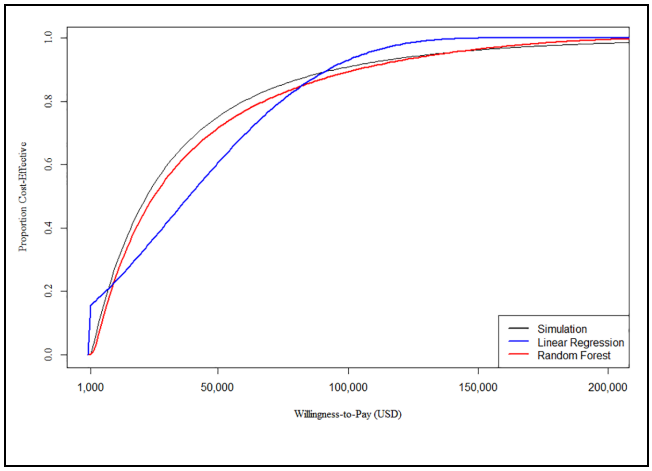

Figure 3 Cost-effectiveness acceptability curve.

The *x*-axis is the ICER values produced by the original simulation model, while the *y*-axis is the ICER values predicted by the LRM (in light gray) and the RFM (in black). The RF metamodel appears both more accurate and precise. The LRM predictions are both more widely spread around the line of perfect fit and are systematically high for low true ICER values and systematically low for high true ICER values.

Figure 3 compares the CEAC produced using the original PSA sample output to the CEACs produced using the predicted ICER values from the LRM and RF metamodel. The RF metamodel predictions systematically underpredicted ICER values across WTP thresholds but

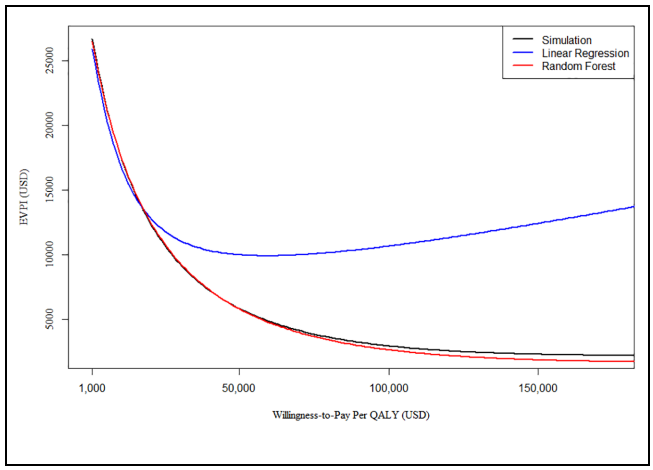

Figure 4 Expected value of perfect information by willingness-to-pay.

remained closer to the CEAC produced using the PSA sample than the LRM predictions for most WTP thresholds. We estimated the average distance between the original and predicted CEACs from WTP thresholds $1,000 to $200,000 to be 0.053 for the LRM and 0.017 for the RF metamodel.

The CEAC was produced by the original simulation for WTP thresholds from $0 to $250,000 using the MATCH simulation model and then recalculated using the 2 metamodels. The solid black line represents the results from the simulation, whereas the dashed blue and red lines are the results from the LRM and RF metamodel, respectively.

Figure 4 shows the performance of our LRM and RF metamodel in predicting EVPI compared with the output from the PSA testing sample. Although both the LRM and RF metamodel accurately predicted EVPI for WTP thresholds below $25,000, the LRM accuracy declined for WTP thresholds beyond this point. The RF metamodel matched the results of the PSA testing sample closely for WTP thresholds up to $200,000. We estimated the average distance between the actual and predicted EVPI curves from WTP thresholds $1,000 to $200,000 to be $8,260 for the LRM and $2,511 for the RF metamodel.

The EVPI was calculated for the WTP thresholds from $0 to $200,000 using the MATCH simulation model and then recalculated using the 2 metamodels. The solid black line shows the results from the simulation, while the dashed blue and red lines are the results from the LRM and RF metamodel, respectively. The RF metamodel outperforms the LRM across all WTP

thresholds and significantly outperforms the LRM for WTP thresholds above $25,000.

Discussion

An ideal metamodel should be accurate, precise, interpretable, and require a reasonably short training time.[8,9] In this study, we built multiple metamodels to replicate the outcomes of a previously validated hepatitis C treatment simulation model that evaluated the cost-effectiveness of hepatitis C treatment. We demonstrated that ML-based metamodels generally outperformed traditional LRMs for cost-effectiveness analysis. The RF metamodel in particular demonstrated a strong ability to mimic simulation model results, both in predicting model outcomes and in conducting CEAC and EVPI. In contrast, the LR metamodel exhibited lower predictive accuracy across all outcomes and showed substantial deviations from simulation results for both CEAC and EVPI. CEAC and EVPI are relatively simple analyses that seldom require a large number of PSA scenarios, so an inability to replicate these results underlines the potential hazards of using LR metamodels for more complex analyses such as EVPPI.

A variety of tools for conducting VOI analyses have been evaluated, including Monte Carlo simulation, efficient calculation with nonparametric regression, and LRMs. Our approach builds on the previous literature by demonstrating the potential of ML-based metamodels to emulate the data-generation process of a complex simulation model more accurately than an LRM without requiring the computation time required for Monte Carlo simulation. In our case, for example, conducting EVPPI with 1000 iterations in both the outer and inner loop (1 million PSA scenarios) would take approximately 1 wk to compute, while the total time required to generate 100,000 PSA scenarios, tune and construct an RF metamodel, and generate predictions for EVPPI would take less than 23 h total (17 for generating the data, 3 for hyperparameter tuning, 12 min for training, and less than 1 min for generating predicted values). As we have demonstrated, an effective ML-based metamodel can generate a large artificial PSA sample in a matter of minutes, and the structure allows researchers to choose which input parameters to vary and which to hold constant when generating metamodel output. Thus, once effectively trained, an ML-based metamodel can be used to conduct multiple VOI analyses in a relatively short period of time.

Published studies in other fields have compared ML-based metamodels among each other and with respect to traditional methods.[14,27–29] Many of these studies find that ML-based methods outperformed LRMs,[27,28] and 1 study we identified similarly found an RF metamodel to perform best.[28] While other studies have compared various metamodeling approaches, this is the first study to our knowledge that performs an extensive comparison of ML-based metamodels on a widely used simulation model with the specific aim of analyzing their application to cost-effectiveness and VOI analysis. In addition to their predictive accuracy, we analyzed each model's ability to replicate EVPI and CEAC, analyses often performed in HEOR. A good fit on these curves is a signal that the fit metamodel can be effectively used for more computationally expensive analyses, such as EVPPI, which often require a large PSA sample for estimation.[11]

A key benefit of ML-based metamodels is that, unlike more complex methods such as GPMs, many easy-to-implement software packages exist for developing ML models that are designed to handle large data sets. Once trained, ML-based models generate predicted values with speed comparable to LRMs. Further, while training GPMs may take days or weeks, ML-based models are often fully trained within hours. Several readily available packages for a wide array of ML-based models exist in R and Python.

Many RF packages, including ranger, have features that can provide additional insights into the original simulation model. Two such features are the impurity measure and the proximity matrix. The impurity measure describes the level of reduction of within-group variance in model outcomes that is achieved, on average, by splitting the training data by a given input parameter.[29,30] For example, if dividing the observations by patient age produces a very large reduction in the within-group variation in QALYs, the impurity measure would regard patient age as a parameter of high importance for determining QALYs. This can allow researchers and policy makers to obtain an unbiased estimate of which parameters are the important drivers of outcomes. The proximity matrix is a measure of how often observations within the original data set fall into the same leaf group of a decision tree and can be viewed as a measure of distance between observations both in terms of their parameters and outcomes. This allows researchers to easily cluster simulations to identify key subgroups within the data.[31] A thorough analysis of these tools is a valuable route for future research.

Generating 100,000 PSA scenarios may be prohibitive for many simulation models. As a supplemental analysis, we trained a new set of metamodels on a random subset of 10,000 PSA scenarios drawn from our original training set, again assessing performance on our full testing set of 900,000. Hyperparameters for ML-based metamodels were once again tuned via a grid search with 5-fold

cross-validation. Even when trained on this smaller set, the testing performance of the RF metamodel remained superior to the traditional LR metamodel. A full table of results can be seen in e-Table 2 in the appendices.

This study has several limitations. A detailed analysis of the RF metamodel revealed that predictions were less precise and accurate for certain patient cohorts. In particular, we found that our model performed worse for young patients beginning in the late stages of hepatitis C, which we attributed to a lack of training data. We next evaluated the performance of both the RF and LR metamodel on a small data set of 10,000 PSA scenarios with the input parameter distributions perturbed. We found the performance of both metamodels declined significantly, illustrating the well-reported dangers of using metamodels to extrapolate to simulation scenarios that deviate too far from the original PSA data-generation process. Another limitation of our study is that although there were correlations between model input parameters in the original simulation, these were ignored when generating PSA samples. An additional limitation of this study is that we did not consider feature selection using established ML-based methods. A robust discussion of feature selection is a valuable route for future research. We did not compare ML metamodels to GPMs, as it was beyond the scope of this study. However, we encourage future research to address this. In particular, we believe that the potential benefit of an ML-based metamodel over a GPM is primarily with regard to the faster training time of the ML algorithm. That is, while this study focuses on the improvements in model performance that accrue from switching to an ML metamodel from a LRM, future research may focus on the potential improvements in training time that accrue from switching to ML metamodel from a GPM. It is important to note that while our findings provide some unique insights, the high performance of our RF metamodels might not be necessarily generalizable. Linear regression or GP metamodels may continue to be preferred for models with a small number of inputs. Further research on the generalizability of ML-based metamodels remains a promising avenue for research.

Conclusion

We found that an ML-based metamodel using RFs outperforms a traditional LRM in replicating the cost-effectiveness analysis of a complex simulation model. In addition to outperforming traditional methods, RF algorithms are a highly intuitive and computationally affordable means of constructing metamodels. Additional benefits of RF models such as accurate measures of variable importance and mechanisms for clustering similar patients provide further added value for researchers and policy makers.[32] RF algorithms could be considered a valuable tool for metamodeling, whether the aim is model replication, cost-effectiveness analysis, the VOI analysis, or developing online decision support tools.

Acknowledgments

We thank Melody Xuan, Maximilliano Cubillos, Mengsha Sun, and Luffina Huang of Georgia Institute of Technology for helping develop many of the ideas used in this study and Yueran Zhuo of Harvard Medical School and Mississippi State University for providing data for developing machine learning models. Data; analytic methods, including code; and materials are available upon request from Jagpreet Chhatwal.

ORCID iD

Jagpreet Chhatwal https://orcid.org/0000-0001-8741-4430

Supplemental Material

Supplementary material for this article is available on the *Medical Decision Making* website at http://journals.sagepub.com/home/mdm.

References

1. Roberts MS, Smith KJ, Chhatwal J. Mathematical modeling. In: Gatsonis C, Morton SC, eds. *Methods in Comparative Effectiveness Research*. Boca Raton (FL): CRC Press; 2017:409–46.
2. Siebert U, Alagoz O, Bayoumi AM, et al. State-transition modeling: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force. *Med Decis Making*. 2012;32(5):690–700. DOI: 10.1177/0272989x12455463
3. Briggs AH, Weinstein MC, Fenwick EA, Karnon J, Sculpher MJ, Paltiel AD. Model parameter estimation and uncertainty: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force—6. *Value Health*. 2012;15(6):835–42. DOI: 10.1016/j.jval.2012.04.014
4. Myers E, Sanders GD, Ravi D, et al. Evaluating the potential use of modeling and value-of-information analysis for future research prioritization within the evidence-based practice center program. Agency for Healthcare Research and Quality (US); June 2011 (Methods Future Research Needs Reports, No. 5). Available from: http://wwwncbinlmnihgov/books/NBK62134/
5. Karnon J. Planning the efficient allocation of research funds: an adapted application of a non-parametric Bayesian value of information analysis. *Health Policy*. 2002; 61(3):329–47.
6. Harris J, Felix L, Miners A, et al. Adaptive e-learning to improve dietary behaviour: a systematic review and cost-

effectiveness analysis. *Health Technol Assess*. 2011;15(37): 1–160.
7. Brush J, Boyd K, Chappell F, et al. The value of FDG positron emission tomography/computerised tomography (PET/CT) in pre-operative staging of colorectal cancer: a systematic review and economic evaluation. *Health Technol Assess*. 2011;15(35):1–192.
8. Jalal H, Dowd B, Sainfort F, Kuntz KM. Linear regression metamodeling as a tool to summarize and present simulation model results. *Med Decis Making*. 2013;33(7): 880–90. DOI: 10.1177/0272989x13492014
9. Tappenden P, Chilcott JB, Eggington S, Oakley J, McCabe C. Methods for expected value of information analysis in complex health economic models: developments on the health economics of interferon-beta and glatiramer acetate for multiple sclerosis. *Health Technol Asses*. 2004; 8(27):1–78.
10. Deegan J Jr. The consequences of model misspecification in regression analysis. *Multivariate Behav Res*. 1976;11(2): 237–48. DOI: 10.1207/s15327906mbr1102_9
11. Rojnik K, Naversnik K. Gaussian process metamodeling in bayesian value of information analysis: a case of the complex health economic model for breast cancer screening. *Value Health*. 2008;11(2):240–50. DOI: 10.1111/j.1524-4733.2007.00244.x
12. Koffijberg H, Degeling K, Ijzerman MJ, Coupé VMH, Greuter MJE. Using metamodeling to identify the optimal strategy for colorectal cancer screening. *Value Health*. 2021;24(2):206–15. DOI: 10.1016/j.jval.2020.08.2099
13. Alam FM, McNaught KR, Ringrose TJ. An artificial neural network based metamodel for analysing a stochastic combat simulation. *Int J Enterp Inf Sys*. 2006;2(4):38–57. DOI: 10.4018/jeis.2006100103
14. De la Fuente R, Smith R. Metamodeling a system dynamics model: a contemporary comparison of methods. *Wint Simul C Proc*. 2017:1926–37.
15. Sabuncuoglu I, Touhami S. Simulation metamodelling with neural networks: an experimental investigation. *Int J Prod Res*. 2002;40(11):2483–505. DOI: 10.1080/00207 540210135596
16. Soeteman DI, Resch SC, Jalal H, et al. Developing and validating metamodels of a microsimulation model of infant HIV testing and screening strategies used in a decision support tool for health policy makers. *MDM Policy Pract*. 2020;5(1):2381468320932894. DOI: 10.1177/2381 468320932894
17. Degeling K, MJ IJzerman, Koffijberg H. A scoping review of metamodeling applications and opportunities for advanced health economic analyses. *Expert Rev Pharmacoecon Outcomes Res*. 2019;19(2):181–7. DOI: 10.1080/ 14737167.2019.1548279
18. Chhatwal J, Kanwal F, Roberts MS, Dunn MA. Cost-effectiveness and budget impact of hepatitis C virus treatment with sofosbuvir and ledipasvir in the United States. *Ann Intern Med*. 2015;162(6):397–406.
19. Aggarwal R, Chen Q, Goel A, et al. Cost-effectiveness of hepatitis C treatment using generic direct-acting antivirals available in India. *PloS One*. 2017;12(5):e0176503. DOI: 10.1371/journal.pone.0176503
20. Chhatwal J, Chen Q, Bethea E, et al. Hep C Calculator: an online tool for cost-effectiveness analysis of DAAs. *Lancet Gastroenterol Hepatol*. 2018;3(12):819. DOI: https://doi .org/10.1016/S2468-1253(18)30281-4
21. Hastie T, Tibshirani R, Friedman JH. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer; 2001.
22. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna (Austria): R Foundation for Statistical Computing; 2020.
23. Lazic SE. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. New York: Springer; 2009.
24. Wright MN, Ziegler A. ranger: A fast implementation of random forests for high dimensional data in C plus plus and R. *J Stat Softw*. 2017;77(1):1–17. DOI: 10.18637/ jss.v077.i01
25. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in Python. *J Machine Learn Res*. 2011;12:2825–30.
26. World Health Organization. *Making Choices in Health: WHO Guide to Cost-Effectiveness Analysis*. Tan-Torres Edejer T, Baltussen R, Adam T, Hutubessy R, Acharya A, Evans DB, Murray DB, Murray CJL, eds. Geneva: World Health Organization; 2003.
27. Bacon B, Gordon S, Lawitz E, Marcellin P, Vierling J, Zeuzem S. HCV RESPOND-2 final results: high sustained virologic response among genotype 1 previous non-responders and relapsers to peginterferon/ribavirin when re-treated with boceprevir plus pegintron (peginterferon alfa-2b)/ribavirin. *Hepatology*. 2010;52.
28. Østergård T, Jensen RL, Maagaard SE. A comparison of six metamodeling techniques applied to building performance simulations. *Appl Energy*. 2018;211:89–103. DOI: https://doi.org/10.1016/j.apenergy.2017.10.102
29. Villa-Vialaneix N, Follador M, Ratto M, Leip A. A comparison of eight metamodeling techniques for the simulation of N2O fluxes and N leaching from corn crops. *Environ Model Softw*. 2012;34:51–66. DOI: https://doi .org/10.1016/j.envsoft.2011.05.003
30. Altmann A, Tolosi L, Sander O, Lengauer T. Permutation importance: a corrected feature importance measure. *Bioinformatics*. 2010;26(10):1340–7. DOI: 10.1093/bioinfor matics/btq134
31. Strobl C, Boulesteix AL, Kneib T, Augustin T, Zeileis A. Conditional variable importance for random forests. *BMI Bioinform*. 2008;9:Article 307. DOI: 10.1186/1471-2105-9-307
32. Alhusain L, Hafez AM. Cluster ensemble based on random forests for genetic data. *Biodata Min*. 2017;10:Article 37. DOI: 10.1186/s13040-017-0156-2