

Machine Learning for Sub-Population Assessment: Evaluating the C-Section Rate of Different Physician Practices

Rich Caruana PhD¹, Radu S. Niculescu², R. Bharat Rao PhD³, Cynthia Simms MD⁴

¹Cornell University, Computer Science, Ithaca, NY

²Carnegie Mellon University, Computer Science, Pittsburgh, PA

³Siemens Corporate Research, Princeton, NJ 08540

⁴Department of Obstetrics, Magee-Womens Hospital, Pittsburgh PA

ABSTRACT

We apply machine learning to the problem of subpopulation assessment for Caesarian Section. In subpopulation assessment, we are interested in making predictions not for a single patient, but for groups of patients. Typically, in any large population, different subpopulations will have different "outcome" rates. In our example, the C-section rate of a population of 22,176 expectant mothers is 16.8%; yet, the 17 physician groups that serve this population have vastly different group C-section rates, ranging from 11% to 23%. The ultimate goal of subpopulation assessment is to determine if these variations in the observed rates can be attributed to (a) variations in intrinsic risk of the patient sub-populations (i.e. some groups contain more "high-risk C-section" patients), or (b) differences in physician practice (i.e. some groups do more C-sections). Our results indicate that although there is some variation in intrinsic risk, there is also much variation in physician practice.

1. INTRODUCTION

In subpopulation assessment, the goal is to make accurate predictions for sets of cases, not for a single case. We would like to assess if different groups have different intrinsic rates for a predicted outcome, or if the variation in outcomes observed across groups is due to extrinsic factors.

Intrinsic factors are those given as inputs to models that are trained to predict the outcome. Extrinsic factors are all factors not entailed by the model's inputs. The basic idea is that the model is allowed to use intrinsic variables to model cases. If the model is accurate, it will correctly compensate for differences between patients (or groups of patients) caused by the intrinsic variables, but will not compensate for differences due to the extrinsic variables.

Our domain is the prediction of Caesarian Section. Our goal is to assess if different groups of patients have different intrinsic risks for C-Section. Note that we are not trying to predict C-Section for any one patient. We want to make predictions of the aggregate intrinsic risk for C-Section in sub-groups of the population. In our example, we observe that the C-section rate of a population of 22,000 expectant mothers is 16.8%; yet, the 17 physician groups that serve this population have vastly different C-section rates, ranging from 11% to 23%. So one question we would like to answer is why the group with 23% C-

section rate is doing more C-sections; is this higher rate in fact warranted by the subpopulation they service, or is it because the physicians in this practice are more likely to use C-Sections than other physicians? The ultimate goal of sub-population assessment is to determine if these variations in observed rates can be attributed to variations in *intrinsic risk* of the patient sub-populations (i.e., some groups see more "high-risk C section" patients), or they are due to differences in physician practice (i.e. some groups do C-sections more often).

In Section 2 we briefly introduce the problem of C-Section rate. A chi-square test was used to show that the variation in the observed C-section rates is in fact significant. We use Leave-One-Group-Out (LOGO) Cross Validation, a modification to traditional cross validation to show that no subgroup of patients contains a disjoint portion of patient population (e.g., all the expectant mothers with diabetes do not all belong to only group 1).

In Section 4 we predict the intrinsic risk of each subpopulation, and assign the difference between the predicted intrinsic population risk and observed risk, to physician practice. This is suggested as a means of ranking physician groups based upon true (i.e. patient-risk adjusted) performance. We examined the performance of the models on the different subgroups to show that much of this difference can in fact be assigned to physician practice, and is not due to poor performance of our models on some of the patient groups.

Section 5 points out several differences between machine learning for subpopulation assessment and for making predictions for single patients.

2. BACKGROUND

2.1 Problem Definition

In the U.S., about 17% of births are by C-section. In Europe, the C-section rate is substantially lower, but patient outcomes do not appear to be worse. Lowering the rate of c-section in the U.S. has both medical and financial benefits. Insurance companies in the U.S. have begun applying financial pressure to lower the c-section rate. One such policy is to pay for a fixed percentage of c-sections, and the practice must make-up the difference if they have a rate higher than the reimbursed rate. If a practice has a rate lower than the quota, it makes more profit.

There are problems with using financial pressure to lower c-section rate. One problem is the tragedy of the commons: individual doctors often have incentives not to lower their c-section rate, even though *groups* of physicians would benefit by lowering their group rate. This problem is complicated by the fact that not all doctors see patients of equal risk. Some doctors specialize in high risk pregnancies and thus should have a higher c-section rate. But how much higher? To evaluate practices fairly, an objective model needs to be developed that can predict whether or not patients *should have* received c-section.

2.2 Magee c-section Database

The database is from Magee-Women's Hospital. It contains 22,176 patients from 1995-1997. Each record has 144 attributes. Previous research [2] left the data clean and ready for this study. Each patient in the database is from one of 17 different physician group practices. The specific goal of this research is to identify physician groups for which the actual c-section rate and the rate predicted by our model differ significantly.

2.3 Prior Work

Dr. Poma notes in [4] that the rate of c-section in the U.S. increased significantly, yet there has not been a related improvement in neonatal outcomes, suggesting that the C-section rate is unnecessarily high.

A report [3] by the Pennsylvania Health Care Cost Containment Council states that cesarean deliveries carry increased risk of complications and longer patient recovery times as well as higher health care costs. The average cost of a c-section in Southwestern PA in 1998 was \$7,885 and the average cost for a vaginal delivery was \$4,787. There are benefits to a lower c-section rate if outcomes are not adversely affected.

In [3], an effort is made to use commonly agreed upon c-section risk factors to distinguish between high risk and low risk patients. In [4], an attempt is made to determine characteristics of obstetricians that affect an obstetrician's c-section rate. Links are drawn correlating the following factors to lower C-section rates: younger obstetrician age, graduation from a domestic medical school, belonging to a group practice, and a smaller number of births.

The authors of [1] compared the c-section rates of different hospitals, but needed to correct for the fact that each hospital saw a different patient population, and thus had patients with different risks for c-section. They constructed a logistic regression model to predict patient risk. Recently, studies by members of our group indicated that machine learning methods (decision trees and artificial neural nets) might be preferable to logistic regression for this problem [2].

3. APPROACH

3.1 Decision Trees

We performed preliminary experiments with decision trees and artificial neural nets. The results of these experiments are

summarized in Table 1. We were also interested in the resubstitution performance because in a retrospective analysis we will be making predictions for patients which we have already seen, and resubstitution yields a more conservative prediction. See Section 5 for a discussion of why the conservative resubstitution prediction is to be preferred in subpopulation assessment.

Previous work with this data [2] used Buntine's IND decision tree software[6]. The results in Table 1 suggest that best performance is obtained with IND's MML decision tree type. MML decision trees are grown to full size (often many thousands of nodes on large datasets with many attributes) and are not pruned. Instead, Bayesian smoothing is applied to the tree to yield predictions at leaf nodes that are a function of the class probabilities along the entire path leading to each leaf node. We often find that MML trees excel at predicting probabilities for problems with discrete outcomes, and thus often perform well for ROC analysis.

Three types of analysis were done with decision trees: 10-fold cross-validation, leave-one-group-out (LOGO) and resubstitution. 10-fold CV was used to determine the generalization performance of the trees. The LOGO and resub analyses were used to predict the C-section rate for each physician group.

In order to do Leave-one-group-out, the data was split into 17 groups based on the physician group.

1. A tree was trained on all groups except group i (i.e. groups 2-16.)
2. Group i was evaluated using the tree i.e. for each patient, the tree was used to determine the probability she would need a C-section.
3. The probabilities were averaged to determine the expected C-section rate for the physician group i .

During LOGO analysis, the model was checked for calibration by verifying that the predicted c-section rate on the training set was approximately equal to the observed c-section rate on the training set. The final result of the LOGO test was an average predicted c-section rate for each of the 17 physician groups. (Results are presented later in Section 4.

3.2 Neural Networks

We also ran the experiments described above using SNNS (Stuttgart Neural Network Simulator) software package. Experiments were conducted to determine the best network architecture and parameter sets. ROC area was used for early stopping. Each network was evaluated by 10-fold CV. Best neural net architectures are also described in Table 1. Both best neural nets performed best when trained with learning rate 0.1. After 10-fold CV was done to select a good set of network parameters (layers, hidden nodes, and learning rate), a leave-one-group-out (LOGO) and resubstitution analysis was performed. The leave-one-group-out procedure is similar to that used for decision trees.

Calibration was a concern for neural networks. When a neural network was tested on its training set, it was found that the predicted C-section rate and the actual C-section rate on the training data differed. A correction factor (the ratio of the actual C-section rate of the training data to the predicted C-section rate of the training data) was applied to the output of the neural network to scale the output into probabilities.

4. RESULTS

Table 1 compares the performance of MML and SMML decision trees, and two different neural net architectures. The performance is measured two ways. The Test ROC column is the average ROC from a 10-fold cross validation. The resubstitution ROC is the ROC computed over the data on which the models were trained. As expected, the resubstitution ROCs are larger than the 10CV ROCs. The resubstitution ROCs for MML trees shows more overfitting than the other models. This is because MML trees tend to be very large, and thus often overfit the data. Note, however, that this overfitting is not detrimental to the 10CV performance of the MML trees.

Figure 1 presents a comparison of Neural Nets vs MML decision trees for predicting group risk.

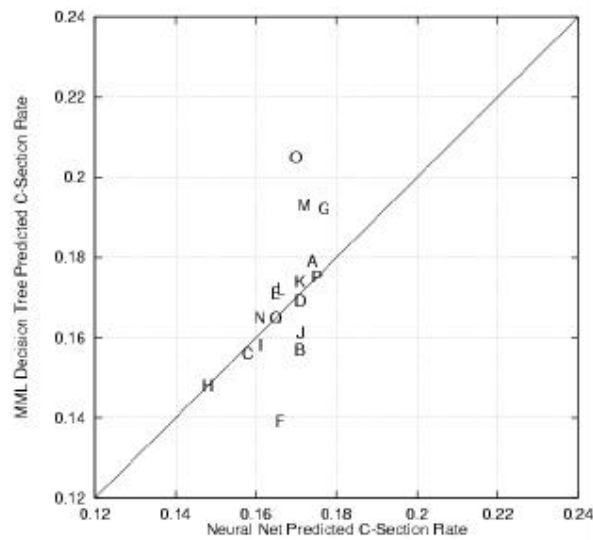


Figure 2 is a scatter plot of the observed c-section rate vs. the predicted rate for each of the 17 physician groups. This figure is drawn to the same scale as Figure 1 to facilitate comparison. The predicted rate is the average of the rates predicted by neural nets, and decision trees, each

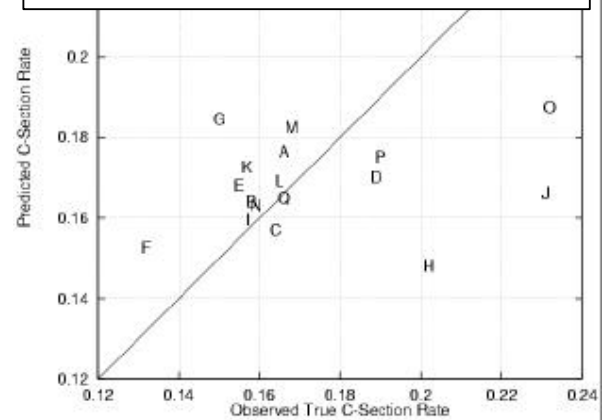
4.1 Subpopulation Assessment

We performed subpopulation assessment using best neural net architecture and MML decision trees. We cannot present the numerical results as a table because of space limitations.

For all models (10CV, LOGO, Resubstitution), we computed the predicted rate for each physician group, and the difference between the observed rate and the predicted rates. The predicted rates are computed by averaging the predictions the models make for each of the patients in a group. Averaging the probabilities yields the total expected number of cesarean sections that each group should have received according to the model. We observed small differences between the 17 physician group rates predicted by Leave One Group Out model and the ones predicted by resubstitution model. This fact points to a low variance of the predictions viewed as random variables, which depends on the particular model chosen.

Model	Test ROC	Resub ROC
ANN 7 hidden units	0.909	0.927
ANN 53 hidden units	0.906	0.93
MML Trees	0.91	0.958
SMML Trees	0.908	0.915

Figure 1: Model ROC Area



trained with both LOGO and resubstitution. Points that fall near the diagonal have an observed c-section rate similar to that predicted by the models. Physician groups having lower c-section rates than the models predict fall in the upper left. Physician groups having c-section rates higher than the models predict fall in the lower right.

Most physician groups fall near the diagonal, indicating that their c-section rates are comparable to the rates predicted by the models. Physician groups H, J, and O, however, exhibit high c-section rates that may not be warranted. Physician group G exhibits a surprisingly low c-section rate. (Group G may be the one group in the study currently practicing at rates comparable to those in Europe). It is important to note that these results to

not necessarily imply that any physician group is doing the wrong thing. These results do, however, point to interesting groups for further investigation.)

The agreement between the neural net and decision tree models (Figure 1) is better than the agreement between the models and the observed rates (Figure 2). In particular, there are no places where the models disagree with each other as much as the models disagree with the observed rates. Interestingly, the group where the models disagree with each other most, group O, is a group for which the models disagree considerably with the observed rate.

5. DISCUSSION

It might seem that subpopulation assessment is simply a matter of applying machine learning to individual cases, and then aggregating the case predictions. This is not true. There are significant differences between learning intended to make predictions for individual cases, and learning for subpopulation assessment.

5.1 Assumptions

Our approach makes several assumptions. One is that by giving models intrinsic variables as inputs, they will be accurate enough to compensate for these factors, yet unable to compensate for extrinsic factors not given as inputs. The difference between the observed and predicted risk is equal to the risk due to these extrinsic factors. These assumptions are not fully justified because we may not capture all variables that relate to the health of the patient (missing inputs), because some of the extrinsic factors may correlate with the intrinsic variables (possibly allowing the model to partially account for extrinsic factors), and because the learned model probably does not learn to make as accurate predictions for cases in the low and high risk tails of the distribution as it does for cases near average risk.

Leave-one-group-out (LOGO) cross-validation makes an assumption that leaving any one group out of the training set does not eliminate an entire class of patient. For example, if *all* patients with diabetes were in group 5, when group 5 was held out the model would not learn about pregnancy with diabetes, and thus might make poor predictions when tested on group 5.

5.2 Overfitting

In traditional machine learning where the goal is to make predictions for individual future cases, overfitting is to be avoided because it reduces generalization performance. In subpopulation assessment, the situation is different. Consider a model with extreme underfitting, e.g., a model that just predicts the average c-section rate for all patients. This is an acceptable, though presumably suboptimal, model for making predictions about individual patients. For subpopulation assessment, however, it is bad. Using it for subpopulation assessment is equivalent to comparing the average rate of each group to the average rate of all groups. Using this model we are most likely to determine that each group is being treated inappropriately, a judgement we should not make unless we are very confident of it.

Now consider a model with extreme overfitting. If we use this model for resubstitution, it predicts that each patient should receive exactly the treatment they received (assuming the database is consistent). When used for subpopulation assessment, this model will assess that each group received the correct c-section rate. Although the overfit model is not sensitive, it is conservative. Thus underfitting leads to overly sensitive tests, and overfitting leads to conservative tests, for subpopulation assessment. In traditional machine learning one usually prefers simpler models that perform well. In subpopulation assessment, we may prefer more complex, somewhat overfit models because they yield more conservative tests.

5.3 Safer Application of Machine Learning

One difference between making predictions for individual patients and for groups of patients is that predictions made for individual patients often are intended to affect the health care provided to the individual. Subpopulation assessment, however, often is used for retrospective analyses. The model will not be used to make care decisions for individual cases---the doctors are still in control. Instead, the model is used to provide feedback to doctors about how their aggregate behavior differs from standard practice.

There is risk associated with providing inaccurate feedback to doctors, but this risk is not as acute as that of making predictions (which may be high variance or inaccurate for some patient subclasses) that decide the health care for an individual. *Subpopulation assessment allows us the freedom to apply advanced machine learning techniques to critical domains in health care without assuming the risk that usually prevents the models from being used in practice.*

5.4 Calibration vs Relative Accuracy

In per-case prediction, usually relative prediction accuracy is more important than absolute accuracy. This is one reason why ROC analysis usually is used in medical decision making -- it is more important to get the ordering of patients by risk correct than it is to get the absolute prediction correct for any one patient. This is not the case with subpopulation assessment. In subpopulation assessment, calibration is more important than relative risk, because it is the aggregate risk that counts, and errors in calibration probably will not be corrected by group aggregation. The models might have perfect ROC, but if their calibration is off, the differences between the group predictions and the observed rates will be wrong.

6. FUTURE WORK

The methods we used do not easily yield confidences in their predictions. Methods that provide confidences would allow us to generate confidence intervals for each group's predicted c-section rate. This, combined with the observed rates and the number of patients in the groups might allow us to perform statistical tests on the difference between the observed and predicted rates.

Best evidence that c-section rate can be lowered without adversely affecting outcomes comes from other countries that have lower cesarean rates but comparable outcomes. A second database containing patients from a such a country might allow us to perform analyses not possible with the U.S. database alone.

We applied subpopulation assessment to the problem of predicting C-Section Rate in different groups of patients. Subpopulation assessment is applicable to many problems in medical decision making. For example, we might want to determine if different subpopulations of patients with heart disease receive different rates of coronary bypass because they have different risk, or because of other factors such as patient socio-economic group, care provider (e.g., small practice vs. large practice, or specialty practice vs. general practice), or health care insurance (e.g., HMO vs. PPO (pay-per-use)).

7. SUMMARY

We used decision trees with Bayesian smoothing and artificial neural nets for *subpopulation assessment*. In subpopulation assessment, the goal is not to make accurate predictions for a single case, but to make accurate *aggregate* predictions for *groups* of cases.

Our analysis indicates that one group of patients who had a c-section rate of 23% (6% higher than the population average) probably does not have enough elevated risk to warrant this c-section rate. Other factors not included in the learned models such as patient and physician preferences or health care funding might explain why this group received more cesareans. Most patient groups had predicted c-section rates similar to the observed rates.

REFERENCES

- [1] Bailit JL, Dooley SL, Peaceman AN. Risk Adjustment for Interhospital Comparison of Primary Cesarean Rates. J. Obstetrics and Gynecology 1999; 93:1025-1030.
- [2] Sims CJ, Myen L, Caruana R, Rao RB, Mitchell T, Krohn M. Predicting cesarean delivery with decision tree models. J. Obstetrics and Gynecology, 2000.
- [3] Pennsylvania Health Care Cost Containment Council. C-Section and Vaginal Deliveries In Southwestern Pennsylvania. Report dated July 1999.
- [4] Poma PA. Effects of obstetrician characteristics on cesarean delivery rates: A community hospital experience. J. Obstetrics and Gynecology 1999; 180:1364-1372.
- [5] Provost F, Fawcett T, Kohavi R. The Case Against Accuracy Estimation for Comparing Induction Algorithms. Proceedings of the Fifteenth International Conference on Machine Learning.
- [6] Buntine W. IND Software for Generating Decision Trees. Available from: URL: [http://ic-](http://ic-
www.arc.nasa.gov/ic/projects/bayes-group/ind/IND-program.html)

[www.arc.nasa.gov/ic/projects/bayes-group/ind/IND-program.html](http://ic-
www.arc.nasa.gov/ic/projects/bayes-group/ind/IND-program.html).

- [7] Caruana R. An Non-Parametric EM-Style Algorithm for Imputing Missing Values. Proceedings of Artificial Intelligence and Statistics 2001, T. Jaakkola and T. Richardson, eds.
- [8] Universities of Stuttgart and Tubingen. SNNS: Stuttgart Neural Network Simulator. Available from: URL: <http://www-ra.informatik.uni-tuebingen.de/SNNS/>.