Prediction of Hospitalization Cost for Childbirth

Si-Chi Chin, James Marquardt, Rui Liu, Martine De Cock*
Center for Data Science
Institute of Technology
University of Washington-Tacoma
{scchin,jamarq,ruiliu,mdecock}@uw.edu

ABSTRACT

Improving cost transparency in healthcare creates opportunities to empower patients and reduce spending. Medical prices vary widely within U.S. markets but few consumers (i.e., patients) have the information to shop for their healthcare services. Consumers are starting to question why their healthcare bills are so high – and why they can't find healthcare prices at all. In this work we apply predictive analytics to help consumers make educated decisions related to childbirth cost. Regression decision trees are used to represent the path from the patients' characteristics (e.g. demographics, comorbidities), choices of medical procedures, and choices of care providers, to the predicted cost variations. The outcome of this research helps consumers make informed choices as well as reduce healthcare expenditures. We demonstrate a system for childbirth cost prediction that far outperforms our baseline model in terms of prediction error. Additionally, we identify several attributes, such as labor room usage, operating room services, anesthesia, and hospital cost to charge ratio, which play a significant role in determining the overall cost of childbirth.

Categories and Subject Descriptors

J.3 [**Computer Applications**]: Life and Medical Sciences—*Health Informatics*

1. INTRODUCTION

According to the Centers for Medicare & Medicaid Services (CMS) National Health Expenditure Accounts (NHEA) reports, in 2012 U.S. healthcare spending increased 3.7 percent to reach $2.8 trillion, or $8,915 per person[1]. These same reports also indicate that healthcare costs for the same procedure vary widely. Childbirth is among the most common reasons for hospitalization in the U.S., accounting for $16.1 billion in hospital costs in 2008 [14]. Recent research [6] has found that, after adjusting for patient demographic and clinical characteristics, the average California woman could be charged as little as $3,296 or as much as $37,227 for a vaginal delivery, and $8,312-$70,908 for a caesarean section depending on which hospital she was admitted to. However, very few consumers (i.e., patients) have the information to shop for their healthcare services. Consumers are starting to question why their healthcare bills are so high – and why they can't find healthcare prices at all. Our work aims to enhance cost transparency and facilitate the decision making process of healthcare consumers.

Hospital charges can affect specific healthcare consumers profoundly. At a time when high-deductible insurance products (e.g. health saving accounts) are increasingly publicized and encouraged by employers, consumers are undergoing growing pressure to make cost-efficient healthcare decisions. However, complicated and opaque healthcare pricing systems have posed immeasurable challenges for consumers to ration their consumption based on their own needs and preferences [11]. The transparency of healthcare pricing is increasingly concerning.

In this work, we use data mining techniques to improve the transparency of hospitalization costs for childbirth. We focus on predicting hospital charges for childbirth and analyzing the contributing factors of the costs. Our research question is formalized as a supervised learning problem. We apply linear regression models as well as regression trees to predict costs of childbirth using State Inpatient Databases (SID) of the State of Washington provided by the Healthcare Cost and Utilization Project (HCUP). In addition, we describe and visualize the effect of adding factors as they become known to consumers towards cost predictions. Last, we apply similar models to predict length of stay of each hospitalization – the leading contributing factor to hospital charges.

Our work is the first, to our knowledge, to apply data mining methods in predicting costs of hospitalization for childbirth. This research also sheds light on developing health consumer decision support tools to navigate the complexities of care options and cost estimates. The contributions of this paper are trifold:

- First, we propose using regression trees to predict childbirth costs in contrast with linear regression models, unveiling the cost factors at different stages of the decision path and allowing making predictions using only partial information available to consumers;
- Second, we categorize attributes according to their availability during a hospitalization and analyze the effect

*On leave from Ghent University
[1]http://www.cms.gov

of each factor set on the prediction quality, enhancing cost transparency for both healthcare providers and consumers;

- Third, we analyze and predict length of stay, the leading factor for childbirth hospitalization cost, providing clinical insights on methods to reduce length of stay.

The rest of the paper is organized as follows. Section 2 summarizes the selected predictive models. Section 3 describes the State Inpatient Databases and the process of extracting a childbirth cohort from the data, as well as the experimental setup. Section 4 gives details on our experimental results. Section 5 summarizes prior studies relevant to healthcare cost predictions and the applications of regression trees for problems in healthcare. Section 6 concludes our work and identifies future research directions.

2. PREDICTIVE MODELS

Multiple Linear Regression. Linear Regression is among the most common approaches to predict a scalar dependent variable y based on a p-dimensional vector of explanatory attributes $\mathbf{x} = (x_1, x_2, \ldots, x_p)$. In linear regression, data is modeled using linear functions, and unknown model parameters are estimated from the data. All categorical variables are encoded as dummy variables. Given a training dataset with n points $\{y^i, (x_1^i, \ldots, x_p^i)\}$, $i = 1, \ldots, n$, a multiple linear regression model takes the form:

$$y^i = \beta_1 x_1^i + \cdots + \beta_p x_p^i + \varepsilon^i = (\mathbf{x}^i)^{\mathrm{T}}\boldsymbol{\beta} + \varepsilon^i, \qquad i = 1, \ldots, n.$$

Linear regression works well in many modeling applications. It does, however, assume that all predictive variables interact linearly with respect to the response variable. Linear regression is also a global model, where there is a single linear predictive formula being applied to the entire data space. When the data has a large number of attributes which interact in complicated, non-linear ways, like in the problem we deal with in this paper, fitting a single linear global model can be difficult. In scenarios where the weight one variable has towards predicting the response variable is determined by the value of a different variable, linear regression can yield inaccurate results.

Regression Trees. Unlike linear regression, regression trees are more robust to non-linear relationships between predictor and single response variables. In order to facilitate this robustness, fitting a regression tree model requires recursive partitioning of the original data in such a way that the resulting groups can be fitted with simple models. Each of the terminal nodes (i.e., leaves) represents a cell of the partition, and a simple model applies to only that cell. In order to make predictions, the fitted model produces a response value as the result of a series of binary decisions made concerning the values of the predictor variables. Regression trees make fast predictions, with the prediction time bounded by the maximum height of the tree. By looking at the trees, it is easy to understand what variables are important in making the predictions. In this paper, we use the Classification and Regression Tree (CART) [3] algorithm to construct regression trees. If a given node has a deviance which is at least some fraction of the root node's deviance (given as the mindev parameter in some CART implementations) it will be partitioned. The deviance of a node consisting of $i = 1...m$ observations of the response variable y and mean μ is defined as the sum of the deviance for all observations in the node, with observation deviance defined as

$$D(\mu; y_i) = (y_i - \mu)^2$$

The algorithm will select the attribute which maximizes the change in deviance of the parent node and the sum of the deviances of the two child nodes. Lower deviance parameters will generally result in trees with a larger number of nodes constructed from a larger set of the available attributes. The time cost of prediction using larger trees is not necessarily prohibitive considering that the number of nodes visited in prediction is guaranteed to be no greater than the height of the tree.

A significant benefit of models built using regression trees is the intuitive manner in which they can be visualized. Multiple linear regression models involving n variables require n dimensions to visualize, which is problematic in scenarios involving healthcare data. Regression trees, on the other hand, can be easily visualized in two dimensions as a tree of binary decisions with respect to predictor attributes, and leaf nodes indicating predicted values. Previous work in both medical [7] and non-medical [4] settings has noted the ease with which tree-based models can be interpreted by individuals with little background knowledge. Regression trees provide health consumers (patients) an intuitive visual representation that facilitates medical decision making.

Baseline. To gauge the effectiveness of our linear regression and regression tree models, we contrast them with simple baseline models. For cost prediction, we set our baseline model to predict the mean μ of all charges in our cohort for all observations, as the average cost is frequently used in government statistical reports[2]. For length of stay prediction, we set our baseline model to predict the most common length of stay in terms of number of days, as the mode is most likely to be sampled.

3. EXPERIMENTAL DESIGN

In this section we describe the data, the methods of attribute selection, and the experimental design.

3.1 Data: Childbirth Cohort

We use the State Inpatient Databases (SID) of Washington State (referred to as SID-WA in the rest of the paper) of year 2010 and 2011. SID are part of the family of databases developed for the Healthcare Cost and Utilization Project (HCUP)[3]. The dataset contains inpatient discharge records from community hospitals in the State of Washington with all-payer, encounter-level information in 2010 and 2011. SID of one year comprises four files that are associated with patients and their encounters in the hospitals. The four files – core file (CORE), charges file (CHGS), diagnosis and procedure groups file (DXPRGRPS), and disease severity measures file (SEVERITY) – provide 596 attributes in total for a single patient encounter. Each inpatient encounter has a unique identifier KEY, that can be used to link records across files. We used KEY to join the CORE, DXPRGRPS, and SEVERITY files and selected 203 attributes that are relevant to the clinical aspect of a patient encounter as shown

[2]For example, the U.S. Agency for Healthcare Research and Quality, HCUPnet, Healthcare Cost and Utilization Project at: http://hcupnet.ahrq.gov/

[3]`http://www.hcup-us.ahrq.gov/sidoverview.jsp`

in Table 2. In order to understand the effect of hospital types, we captured hospital-level characteristics by joining the data with cost-to-charge ratio files using the provided hospital files. To construct the childbirth cohort for the experiments, we extracted from SID-WA those discharges that have Diagnosis Related Group (DRG) codes listed in Table 1.

Table 1: Diagnosis Related Group (DRG) codes for childbirth cohort

DRG	Description
765	Caesarean section w/ complications and comorbidities (CC)/multiple complications and comorbidities (MCC).
766	Caesarean section w/o CC/MCC
767	Vaginal delivery w/ sterilization &/or dilation and curettage (D&C)
768	Vaginal delivery w/o sterilization &/or D&C
774	Vaginal delivery w/ complicating diagnosis
775	Vaginal delivery w/o complicating diagnosis
781	Other antepartum diagnosis w/ medical complications
782	Other antepartum diagnosis w/o medical complications

3.2 Attribute Selection

The original childbirth cohort is highly skewed. Most patients only utilize a few healthcare services (i.e., revenue codes), which introduces zero or missing values for most attributes as shown in Figure 1. The sparseness of the data creates challenges in modeling as information about certain attributes might be missing in the training data but could still be present in the testing set. We therefore selected a subset of attributes to mitigate this issue. Two methods were used for attribute selection. The first method is *Frequency-based (FB) Selection*. A threshold is selected to filter out those attributes that occur least frequently in the data. Based on our empirical studies and the distribution shown in Figure 1, we include only attributes that have more than 10% non-zero or non-missing values. The second method is *Regression-based (RB) Selection*. Simple linear regression is performed between each explanatory attribute and predicted attribute (i.e., hospital charges). We keep attributes that significantly (p-value < 0.05) contribute to the prediction of charges. The results of attribute selection and the number of attributes – categorized in five groups – are shown in Table 2.[4] The attribute selection results show that RB includes more attributes in comorbidities and revenue codes (see Table 2). We use 37 attributes for FB experiments and 129 for RB experiments.

3.3 Evaluation Methods

We built regression trees using the package "tree" [9] in R[5]. We manipulated the deviance threshold parameter (mindev) to control the size of the tree and the number of attributes used. Three values – 0.1, 0.01, 0.001 – were used for the deviance parameter as the impurity measure for the experiments.

We report three evaluation metrics for our experiments: R^2, Mean Absolute Error (MAE), and Root Mean Square Error (RMSE). To be able to compare our results to published studies that used regression models, we report R^2. However, R^2 is an indicator for the explanatory powers of the model, as opposed to predictive powers. We therefore add MAE and RMSE to evaluate our predictive models. MAE is the average absolute difference between the predicted costs and the true costs of childbirth hospitalization. As an example, if we predict the hospitalization cost for a patient to be \$9,000, but the true cost is \$12,000, then the absolute error is $|\$9,000 - \$12,000| = \$3,000$. While MAE provides a more straightforward explanation on the errors, RMSE diagnoses variation in errors. In our study, large errors are particularly undesirable. The RMSE indicates which predictive model is more robust to outliers in the data.

We validated our prediction results using 10 runs of 10-fold cross-validation for both linear regression and regression tree models. These experiments were repeated across multiple subsets of the total collection of attributes, adding one attribute category per experiment according to the order with which they likely to be known by the consumer; the progression of which is denoted in Table 2. This process is carried out for both the frequency-based and regression-based attribute categories as described in the previous subsection. The sequence of the attribute categories is determined based on the availability of the information from a consumer's perspective. We assume that the base attributes (age, race, and length of stay (LOS)) are known to most consumers and the hospital information such as Wage index and hospital spe-

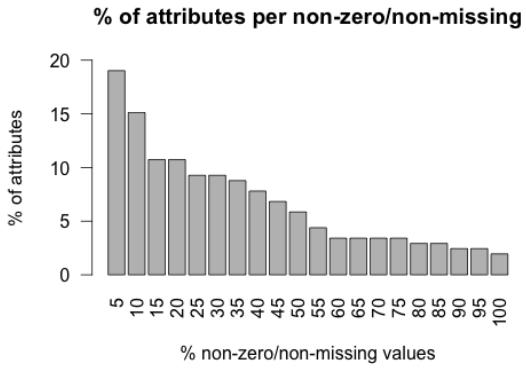


Figure 1: Distribution of non-zero/non-missing attribute values

Table 2: Number of attributes

Attr. Category	All	FB	RB
Base (B)	4	3	3
Comorbidities (C)	29	3	21
Revenue codes (R)	164	25	99
Diagnosis Related Group (D)	1	1	1
Hospital information (H)	5	5	5
Total	203	37	129

[4]A complete listing of variable names, including a look-up table for revenue code interpretations, can be found in `http://cwds.uw.edu/prediction-hospitalization-cost-childbirth-variables`

[5]www.r-project.org/

cific all-payer inpatient cost-charge ratio are least available to consumers.

4. RESULTS AND DISCUSSION

4.1 Decision Path

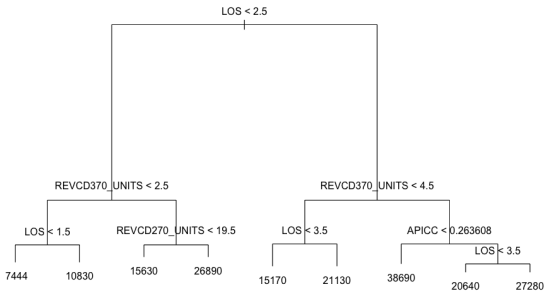


Figure 2: Regression tree visualization example

Figure 2 shows an example regression tree using mindev = 0.01. The numbers at the leaves are the predicted values. The visualization of the regression tree informs patients which attributes have the highest predictive powers for hospitalization costs as well as identifies the interactions between attributes. According to Figure 2, length of stay (LOS) is the leading factor for the costs. Hospitalization costs are likely to be lower if LOS is less than 2.5 days. The next influential factor is the use of anesthesia (REV370). Patients who stay in the hospital for less than 2.5 days may still have high hospital charges if they use more than 2.5 units of anesthesia and utilize more than 19.5 units of medical/surgical supplies. For patients who stay more than 2.5 days in the hospital, the utilization of anesthesia is the next important factor for their hospital charges. If the use of anesthesia is higher than 4.5 units, the all-payer inpatient cost-charge ratio (APICC), which indicates the inflation of medical price for different hospitals, becomes a decision factor. This finding implies that the choice of hospital can affect the final hospital charges.

The visualization of regression trees, compared to most predictive modeling techniques, provides a more intuitive understanding of cost drivers and their interaction to empower healthcare consumers.

4.2 Factor Analysis

Figure 3 illustrates the effect of each attribute category (see Table 2) on prediction quality. We note that comorbidities provide little information on cost prediction. The choice of procedures (i.e., revenue codes) drives the hospitalization costs. Although the improvement is marginal after adding diagnosis information (DRG), hospital information is shown to be influential to cost prediction.

Among the base attributes, we observe that length of stay is consistently the root node for all regression trees, corresponding to the best predictor. As we add attributes to our regression tree models, we find that the second level decision nodes remain dictated by length of stay until the revenue code category of attributes is added. Once these attributes are added, the second level decisions are dictated by operating room service and anesthesia utilization.

Table 3 compares the top five contributing factors for the best performing linear regression model and regression tree model. We observe that in our best performing regression tree models, the five most frequent decision attributes are those containing information on labor room usage, operating room services, anesthesia, and the all-payer inpatient cost-charge ratio of the visited hospital. For linear regression, we observe the five highest weighted attributes after feature normalization to be related to therapeutic services, drug administration, pharmacy utilization, area wage index of the hospital, as well as a geographic adjustment factor determined by HCUP, as shown in Table 3.

The difference between important attributes in the linear regression and regression tree models is very informative for the evaluation of each. Not only does length of stay not appear as a highly weighted factor in linear regression, but three of the five highly weighted attributes do not appear as either attributes near the root of the regression trees or as a frequently used decision attribute. Within our attributes, a high degree of collinearity exists between certain variables such as REVCD250_UNITS (Pharmacy, general classification) and REVCD251_UNITS (Pharmacy, generic drugs). These confounding factors led to the linear regression model being unable to determine the contribution of variables to the response variable value appropriately. This issue would have no effect on our regression tree models, as attributes are chosen according to how well they divide observations within a node, not according to least squares as in linear regression.

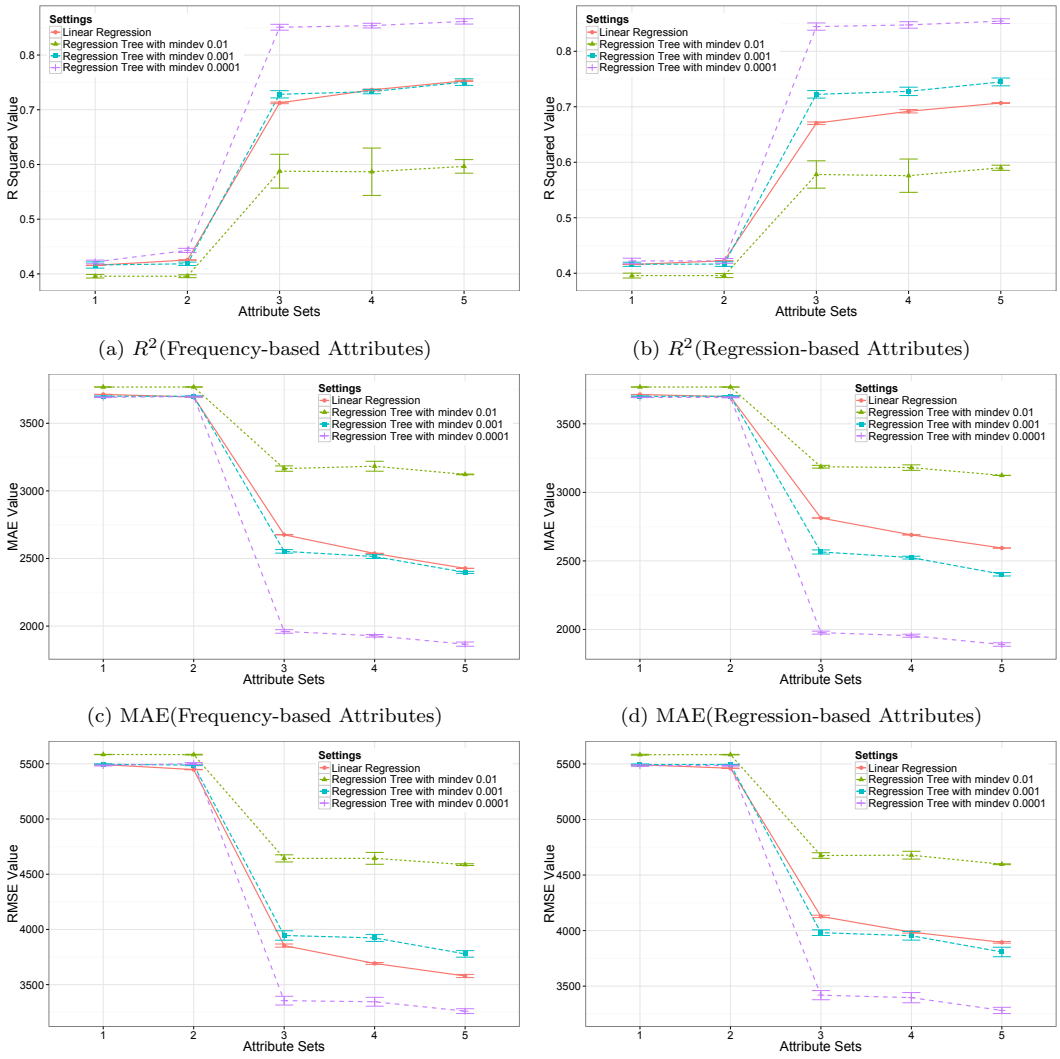
Table 3: Contributors for TOTCHG using the best performing linear regression and CART model. We select the top 5 attributes (sorted in column `Rk`) with highest weights for linear regression and top 5 attributes with highest frequencies in regression tree.

Rk	LR	CART
1	Geographic Adjustment Factor	Labor room/general
2	Area Wage Index	Hospital Cost to Charge Ratio
3	Therapeutic services	Anesthesia
4	Drugs Requiring Detailed Coding	Operating room
5	Pharmacy IV Solutions	Labor room/labor

4.3 R^2, MAE, RMSE

Tables 4-6 present our experimental results for Total Charge. For all experiments across both model approaches, attribute sets, and attribute selection techniques, our models achieve significantly lower RMSE and MAE than those achieved using the baseline model. Recall that the baseline predicts the average cost $\mu = 13132.77$ for all observations.

As seen in Figure 3, the addition of various attribute categories to our models results in different degrees of improvement in error. Initially, all models perform similarly with only attributes from the Base category. The addition of attributes from the Comorbidity category has very little impact on RMSE and MAE for all models. Significant reductions in both error metrics are seen across all models with the introduction of Revenue Code category attributes. The addition of the DRG does not lower the RMSE of our regression tree models, although a small increase in MAE occurs in our model trained with a deviance threshold of 0.01. The addition of the same attribute incurs a perceptible decrease in our linear regression model's RMSE and MAE. Finally,



(a) R^2(Frequency-based Attributes)

(b) R^2(Regression-based Attributes)

(c) MAE(Frequency-based Attributes)

(d) MAE(Regression-based Attributes)

(e) RMSE(Frequency-based Attributes)

(f) RMSE(Regression-based Attributes)

Figure 3: Experimental results. Error bars show 95% confidence interval of 10 runs 10-fold cross-validation results

the addition of attributes from the Hospital category lowers the RMSE and MAE of all models perceptibly.

Linear regression benefits more than regression trees in terms of RMSE and MAE decrease from our regression based feature selection, as seen in Figure 3e. For all versions of our regression tree models, there is no perceptible change in RMSE between features selected using frequency and those selected using regression. Linear regression, on the other hand, sees benefits primarily from the addition of the Revenue Code category of attributes. Table 6 demonstrates that the addition of the Revenue Code subset of the regression selected attributes to the linear regression model improves the decrease of RMSE by 16.3% when compared with the addition of the same category from the frequency selected attributes. A smaller but still significant decrease in MAE of 3.6% is also seen when comparing linear regression models in the same manner.

Regression trees built with a low minimum deviation parameter (mindev = 0.01) result in higher RMSE and MAE than linear regression for all attributes. For frequency selected attributes, regression trees outperform linear regression for all attributes for mindev = 0.001, 0.0001. For regression selected attributes, regression trees outperform linear regression for mindev = 0.0001.

As a third method of evaluating our models, we note improvements in R^2 for all models as attributes are added. As seen in Figure 3a and Figure 3b, the patterns in R^2 increase are similar to those observed in MAE and RMSE decrease. Additionally, our linear regression models see more benefits in terms of R^2 increase than our regression tree models. One notable difference in these patterns is that the confidence intervals for R^2 in our regression tree models with deviance thresholds of 0.0001 is large relative to other models after the addition of the revenue code attributes.

Table 4: R^2 for Total Charge prediction. Attribute definitions given in Table 2

Attribute Set	Attributes	Model	mindev	Frequency-Based Attr Mean R^2	Frequency-Based Attr 95% CI (+/-)	Regression-Based Attr Mean R^2	Regression-Based Attr 95% CI (+/-)
-	-	Baseline	-	NA	-	NA	-
1	B	Linear Regression	-	0.416	7.295e-05	0.416	1.013e-04
2	B-C	Linear Regression	-	0.422	5.762e-05	0.425	4.748e-04
3	B-C-R	Linear Regression	-	0.670	2.265e-03	0.713	1.473e-03
4	B-C-R-D	Linear Regression	-	0.692	3.024e-03	0.736	1.482e-03
5	B-C-R-D-H	Linear Regression	-	0.707	7.310e-04	0.753	1.088e-03
1	B	Regression Tree	0.01	0.396	4.385e-03	0.396	3.285e-03
1	B	Regression Tree	0.001	0.416	3.785e-03	0.416	5.726e-03
1	B	Regression Tree	0.0001	0.422	4.938e-03	0.422	3.316e-03
2	B-C	Regression Tree	0.01	0.396	3.625e-03	0.396	2.657e-03
2	B-C	Regression Tree	0.001	0.417	4.138e-03	0.419	3.079e-03
2	B-C	Regression Tree	0.0001	0.422	4.325e-03	0.443	3.782e-03
3	B-C-R	Regression Tree	0.01	0.578	2.465e-02	0.588	3.092e-02
3	B-C-R	Regression Tree	0.001	0.722	6.787e-03	0.728	6.631e-03
3	B-C-R	Regression Tree	0.0001	0.844	6.558e-03	0.851	5.357e-03
4	B-C-R-D	Regression Tree	0.01	0.576	2.999e-02	0.587	4.325e-02
4	B-C-R-D	Regression Tree	0.001	0.728	7.555e-03	0.733	3.954e-03
4	B-C-R-D	Regression Tree	0.0001	0.847	5.842e-03	0.854	4.241e-03
5	B-C-R-D-H	Regression Tree	0.01	0.590	4.668e-03	0.596	1.239e-02
5	B-C-R-D-H	Regression Tree	0.001	0.745	7.054e-03	0.751	6.088e-03
5	B-C-R-D-H	Regression Tree	0.0001	0.854	4.261e-03	0.861	4.848e-03

Table 5: MAE for Total Charge prediction. Attribute definitions given in Table 2

Attribute Set	Attributes	Model	mindev	Frequency-Based Attr Mean MAE	Frequency-Based Attr 95% CI (+/-)	Regression-Based Attr Mean MAE	Regression-Based Attr 95% CI (+/-)
-	-	Baseline	-	8319.834	-	8319.834	-
1	B	Linear Regression	-	3713.787	0.175	3713.840	0.303
2	B-C	Linear Regression	-	3697.859	0.375	3694.746	0.711
3	B-C-R	Linear Regression	-	2812.898	0.762	2675.603	1.794
4	B-C-R-D	Linear Regression	-	2689.928	0.925	2537.056	1.828
5	B-C-R-D-H	Linear Regression	-	2594.008	1.164	2427.069	0.776
1	B	Regression Tree	0.01	3768.385	0.250	3768.389	0.181
1	B	Regression Tree	0.001	3700.492	2.666	3700.024	2.242
1	B	Regression Tree	0.0001	3692.866	1.595	3692.805	2.570
2	B-C	Regression Tree	0.01	3768.386	0.228	3768.391	0.137
2	B-C	Regression Tree	0.001	3701.254	4.322	3700.021	3.341
2	B-C	Regression Tree	0.0001	3692.768	3.026	3697.663	3.472
3	B-C-R	Regression Tree	0.01	3187.107	10.111	3165.101	19.880
3	B-C-R	Regression Tree	0.001	2565.084	15.078	2552.870	13.170
3	B-C-R	Regression Tree	0.0001	1976.846	10.912	1960.459	13.217
4	B-C-R-D	Regression Tree	0.01	3180.622	20.050	3182.447	36.735
4	B-C-R-D	Regression Tree	0.001	2524.049	10.084	2514.160	13.813
4	B-C-R-D	Regression Tree	0.0001	1954.354	11.069	1928.089	9.765
5	B-C-R-D-H	Regression Tree	0.01	3123.747	0.326	3121.932	2.618
5	B-C-R-D-H	Regression Tree	0.001	2402.799	12.706	2397.349	7.316
5	B-C-R-D-H	Regression Tree	0.0001	1890.143	12.809	1866.316	15.574

In addition to observations made on RMSE, MAE, and R^2 improvements for our various models, we observe the structure of the best performing representatives of both regression trees and linear regression to understand which attributes are most important for cost prediction.

4.4 Length of Stay Analysis

As a result of the observation that length of stay is consistently the root decision attribute of our regression trees, we conduct an ad-hoc experiment to predict length of stay using linear regression and regression trees. Our experimental design is identical to the setup described in Section 3.3, with the addition that leakage variables (e.g. REVCD120_UNITS, which has units in days) are removed from our attribute set. Figure 4 and Table 7 show the effect of each attribute category on prediction of length of stay. The baseline method uses the most frequent length of stay (2.5 days) for the prediction. We observe that linear regression models outperform regression trees when only the basic attributes and comorbidities are available. Revenue codes (i.e., prescribed procedures) largely improve the prediction quality for all models. It is noted that regression trees consistently outperform linear regression on predicting length of stay when revenue codes, DRG, and hospital information are available for making predictions.

As shown in Table 7, we observe significant prediction improvements after adding revenue code attributes. It is also noted that, as shown in Figure 4, regression tree models outperform linear regression after incorporating revenue codes, DRG, and hospital information. While the lists of most influential factors determined by each model differ, it should be noted that they intersect on two variables: general pharmacy utilization and self administered drugs. Aside from these two variables, we see that regression trees also indicate labor room usage for labor, drugs requiring detailed coding, and laboratory usage to be highly important factors. In contrast, our linear regression model gives high

Table 6: RMSE for Total Charge prediction. Attribute definitions given in Table 2

Attribute Set	Attributes	Model	mindev	Frequency-Based Attr Mean RMSE	Frequency-Based Attr 95% CI (+/-)	Regression-Based Attr Mean RMSE	Regression-Based Attr 95% CI (+/-)
-	-	Baseline	-	7186	-	7186	-
1	B	Linear Regression	-	5492.717	0.223	5492.849	0.338
2	B-C	Linear Regression	-	5461.376	0.239	5447.397	1.933
3	B-C-R	Linear Regression	-	4127.723	10.660	3853.617	14.364
4	B-C-R-D	Linear Regression	-	3986.496	9.006	3692.116	8.191
5	B-C-R-D-H	Linear Regression	-	3894.482	7.522	3578.202	14.224
1	B	Regression Tree	0.01	5582.874	1.172	5583.946	0.520
1	B	Regression Tree	0.001	5494.803	1.622	5495.520	1.548
1	B	Regression Tree	0.0001	5484.718	2.008	5484.283	1.543
2	B-C	Regression Tree	0.01	5583.621	0.903	5582.961	1.170
2	B-C	Regression Tree	0.001	5494.797	1.446	5488.329	2.028
2	B-C	Regression Tree	0.0001	5484.336	1.948	5501.603	3.230
3	B-C-R	Regression Tree	0.01	4641.890	19.336	4643.445	11.811
3	B-C-R	Regression Tree	0.001	3958.254	11.957	3945.171	15.632
3	B-C-R	Regression Tree	0.0001	3343.247	19.796	3355.011	14.405
4	B-C-R-D	Regression Tree	0.01	4634.147	16.580	4643.649	19.599
4	B-C-R-D	Regression Tree	0.001	3916.610	15.412	3923.147	11.561
4	B-C-R-D	Regression Tree	0.0001	3332.642	14.269	3345.022	14.656
5	B-C-R-D-H	Regression Tree	0.01	4594.164	3.927	4586.713	2.897
5	B-C-R-D-H	Regression Tree	0.001	3777.487	16.470	3778.316	10.842
5	B-C-R-D-H	Regression Tree	0.0001	3245.585	14.221	3260.104	7.942

weight to geographic adjustment factor, area wage index, and pharmacy IV solution utilization.

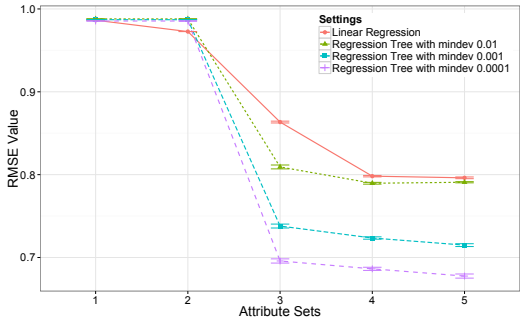


Figure 4: Prediction errors for length of stay

5. RELATED WORK

Childbirth is among the most common reasons for hospitalization in the U.S., accounting for $16.1 billion in hospital costs in 2008 [14]. At a time when the rising cost of healthcare has come to the forefront of political and medical discourse, the transparency of healthcare pricing is increasingly concerning [11, 6]. Therefore, predicting such costs accurately is an indispensible first step in addressing this problem.

Numerous risk-adjustment models are developed by health services researchers, providers, and payers to estimate healthcare cost for each patient and to compare with the observed cost [12, 16]. Most earlier research used classical regression models [8, 15] to predict healthcare costs, or logistic regression [5, 16] to analyze risk factors. However, except in [8], the majority of these regression studies do not perform out-of-example experiments to validate the predictive powers of the models [2]. In our experiments, we perform 10 runs of 10-fold cross validation as evidence of the prediction accuracy of our methods.

More recent research has investigated using data mining methods [2, 13] to predict healthcare costs. Bertsimas et al. [2] utilized classification trees and clustering algorithms to predict annual medical costs for individuals. Srinivasan and Arunasalam [13] utilized statistical and data mining techniques to predict hospital, medical, and prosthetic costs for each individual hospital admission. In contrast with this work, we emphasize the use of regression trees to clarify the origins of medical costs and to visualize the available decision paths. In addition, we systematically analyze the effect of different types of attributes on the predictive power of our methods. Moreover, we concentrate on predicting hospitalization costs for childbirth, using 2 years (2010 and 2011) data from SID-WA provided by HCUP.

Two studies [1, 6] utilized classical linear regression models to analyze hospitalization costs for childbirth. Hsia et al. [6] investigated the between-hospital variation of charges and discounted prices for uncomplicated vaginal and caesarean section deliveries, using data from the California Office of Statewide Health Planning and Development (OSHPD). The authors found that the charges are significantly influenced by institutional and market-level factors, such as wage index, percentage of uninsured individuals in the county, etc.. However, the majority of variations in charges between hospitals remains unexplained. Allen et al. [1] examined the association between the number of breastfeeding supportive practices and costs of an uncomplicated birth, using the 2007 Maternity Practices in Infant Nutrition and Care survey (mPINC) and SID from HCUP. The authors indicated that the number of ideal practices is not a significant attribute for higher childbirth costs. Both studies [1, 6] emphasized the influence of specific factors to hospitalization costs for childbirth. Compared to these studies, we explore a much wider range of attributes, as described in Section 3. In addition, we report lower prediction errors of our data mining methods in contrast with classical linear regression models.

Robinson et al. [10] examined the economic impact of performing elective repeat caesarean during 37 or 38 weeks of gestation relative to 39-week delivery, recommended by the American College of Obstetricians and Gynecologists.

Table 7: RMSE for Length of Stay Prediction. Attribute definitions given in Table 2

Set	Attributes	Model	mindev	Mean RMSE	95% CI (+/-)
	-	Baseline	-	1.017407	-
1	B	Linear Regression	-	0.986	4.192e-05
2	B-C	Linear Regression	-	0.973	5.263e-05
3	B-C-R	Linear Regression	-	0.863	1.101e-03
4	B-C-R-D	Linear Regression	-	0.798	8.657e-04
5	B-C-R-D-H	Linear Regression	-	0.7962	1.036e-03
2	B	Regression Tree	0.0001	0.985	5.460e-05
2	B-C	Regression Tree	0.0001	0.985	9.810e-05
3	B-C-R	Regression Tree	0.0001	0.696	9.247e-4
4	B-C-R-D	Regression Tree	0.0001	0.686	6.955e-4
5	B-C-R-D-H	Regression Tree	0.0001	0.678	9.090e-4

The authors used decision analysis modeling to evaluate economic outcomes for a hypothetical cohort of neonates. The authors used Florida HCUP SID and Eunice Kennedy Shriver National Institute of Child Health and Human Development Maternal-Fetal Medicine Units Network (NICHD-MFMU) data for their work. The total cost of care for each neonate hypothetical cohort was calculated based on the probability of anticipated incidence of the adverse event. Compared to [10], we study both vaginal and caesarean birth, including both complicated and uncomplicated samples. We perform extensive systematic experiments using a CART implementation of regression trees to evaluate the predictive power of our methods.

6. CONCLUSION AND FUTURE WORK

In this work we demonstrated the effectiveness of both linear regression and regression trees (CART) towards the task of predicting childbirth hospitalization costs. Regression tree models outperform linear regression models after incorporating information about healthcare services utilization (revenue codes), the diagnosis information (DRG), and hospital characteristics. Length of stay and the utilization of labor room, anesthesia, and operation room are the drivers for hospital charges for childbirth. The visualization of regression trees (Figure 2) provides patients an intuitive understanding about the leading factors and their interaction at different stages of care. Additionally, as demonstrated in the experiments detailed in Section 4.4, regression trees are particularly powerful in predicting length of stay, one of the attributes shown to be most important in predicting total cost. Investigation into predicting the values of attributes which hospitalized mothers have little control over yet have a large affect on total cost would be valuable.

Future work will explore additional data mining techniques such as clustering methods as well as additional risk factors such as complications and readmission risks to predict hospitalization costs. We will also investigate the effects of confounding factors on linear regression models used to predict costs in the dataset used for this study. Additionally, we will include more historical data to conduct longitudinal studies, incorporating the temporal dimension in predicting hospital charges.

As clinical risks associated with decisions made during the course of care are of primary concern to both care providers and patients, an additional avenue of research we will pursue in future work will be to investigate methods for minimizing both readmission risk and patient cost.

Acknowledgements

This work is supported by MHS (grant no: A73191). Additionally, we are thankful data architects and the clinicians at MHS for their valuable time and insightful discussions during the initial stage of the study, and to Edifecs Inc. for its generous support to the Center for Data Science.

7. REFERENCES

[1] J. A. Allen, H. B. Longenecker, C. G. Perrine, and K. S. Scanlon. Baby-friendly hospital practices and birth costs. *Birth*, 40(4):221–226, 2013.

[2] D. Bertsimas, M. V. Bjarnadóttir, M. A. Kane, J. C. Kryder, R. Pandey, S. Vempala, and G. Wang. Algorithmic prediction of health-care costs. *Operations Research*, 56(6):1382–1392, 2008.

[3] L. Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

[4] S. Chen, A. Borthwick, and V. R. Carvalho. The case for cost-sensitive and easy-to-interpret models in industrial record linkage. In *International Workshop on Quality in Databases VLDB-2011*. Citeseer, 2011.

[5] G. M. Chertow, E. Burdick, M. Honour, J. V. Bonventre, and D. W. Bates. Acute kidney injury, mortality, length of stay, and costs in hospitalized patients. *Journal of the American Society of Nephrology*, 16(11):3365–3370, 2005.

[6] R. Y. Hsia, Y. A. Antwi, and E. Weber. Analysis of variation in charges and prices paid for vaginal and caesarean section births: a cross-sectional study. *BMJ Open*, 4(1):e004017, Jan. 2014.

[7] S. C. Lemon, J. Roy, M. A. Clark, P. D. Friedmann, and W. Rakowski. Classification and regression tree analysis in public health: methodological review and comparison with logistic regression. *Annals of behavioral medicine*, 26(3):172–181, 2003.

[8] C. A. Powers, C. M. Meyer, M. C. Roebuck, and B. Vaziri. Predictive modeling of total healthcare costs using pharmacy claims data: A comparison of alternative econometric cost modeling techniques. *Medical care*, 43(11):1065–1072, 2005.

[9] B. Ripley. *tree: Classification and Regression Trees*, 2014. R package version 1.0-35.

[10] C. J. Robinson, M. S. Villers, D. D. Johnson, and K. N. Simpson. Timing of elective repeat cesarean delivery at term and neonatal outcomes: a cost analysis. *American journal of obstetrics and gynecology*, 202(6):632e1–632e6, 2010.

[11] J. C. Robinson and P. B. Ginsburg. Consumer-driven health care: Promise and performance. *Health Affairs*, 28(2):w272–w281, 2009.

[12] J. W. Robinson. Regression tree boosting to adjust health care cost predictions for diagnostic mix. *Health services research*, 43(2):755–772, 2008.

[13] U. Srinivasan and B. Arunasalam. Leveraging big data analytics to reduce healthcare costs. *IT Professional*, 15(6):21–28, Nov 2013.

[14] A. A. Vashi, J. P. Fox, B. G. Carr, G. D'Onofrio, J. M. Pines, J. S. Ross, and C. P. Gross. Use of hospital-based acute care among patients recently discharged from the hospital. *JAMA*, 309(4):364–371, 2013.

[15] Y. Zhao, A. S. Ash, R. P. Ellis, J. Z. Ayanian, G. C. Pope, B. Bowen, and L. Weyuker. Predicting pharmacy costs and other medical costs using diagnoses and drug claims. *Medical care*, 43(1):34–43, 2005.

[16] Y. Zhao, T.-C. Kuo, S. Weir, M. S. Kramer, and A. S. Ash. Healthcare costs and utilization for medicare beneficiaries with alzheimer's. *BMC health services research*, 8(1):108, 2008.