A Data-Driven approach towards Patient Identification for Telehealth Programs

Martha Ganser*, Sauptik Dhar[†1], Unmesh Kurup[†], Carlos Cunha[††], and Aca Gacic*.

*Robert Bosch Healthcare Systems, Inc., Palo Alto, CA
[†]Robert Bosch Research and Technology Center, Palo Alto, CA
[††]Robert Bosch Data Mining Services, Palo Alto, CA
[1]sauptik.dhar@us.bosch.com

Abstract— **Telehealth provides an opportunity to reduce healthcare costs through remote patient monitoring, but is not appropriate for all individuals. Our goal was to identify the patients for whom telehealth has the greatest impact, as measured through cost savings and patient engagement. For prediction of cost savings, challenges included the high variability of medical costs and the effect of selection bias on the cost difference between intervention patients and controls. Using Medicare claims data, we computed cost savings by comparing each telehealth patient to a group of control patients who had similar healthcare resource utilization. These estimates were then used to train a predictive model using logistic regression. Filtering the patients based on the model resulted in an average cost savings of $10K in the group of patients with the highest healthcare utilization, an improvement over the current expected loss of $2K (without filtering). Groups of patients with lower healthcare utilization also showed improvement, though less pronounced. To identify highly engaged patients, we developed predictive models of telehealth compliance and of patient satisfaction. Performance of these models were generally poor, with an AUC ranging from 0.54 to 0.64.**

Keywords -- healthcare; telehealth; logistic regression.

I. INTRODUCTION

Telehealth can prevent costly healthcare interventions through continuous care with remote monitoring [1]. Patients engage daily by taking vital signs and learning disease self-management skills. Meanwhile, a care team monitors the patients' status daily, assessing risk and providing intervention when necessary. The data generated is stored in a data warehouse where it can be further linked to medical insurance claims that contain information on patients' healthcare utilization, diagnoses, procedures, and costs. Although telehealth enabled care management has proven financial and healthcare benefits, there are costs associated with distributing the monitoring devices and operating the care management staff. In order to maximize the benefits, patients may be prioritized for enrollment according to the expected impact on health outcomes and costs.

Currently, patients are selected for telehealth programs using clinical groupers and statistical models. Clinical groupers provide a method of categorizing patients by the level of healthcare resource use and morbidity [2]. Statistical models have also been developed to predict patients who are high cost or at high risk of hospitalization [3]. Although both methods are useful for identifying high risk patients, not all high risk patients are a good fit for telehealth. For example, patients with end-stage renal disease (ESRD) are high risk but difficult to impact in a way that would reduce the cost for a payer. Conversely, patients with lower risk scores may benefit from preventative measures. Intervention-specific models will allow us to select the most appropriate patients for telehealth.

In addition to having the right medical characteristics, it is important to enroll patients that will actively engage with the telehealth system. Broadly speaking, patient engagement refers to an individual's level of involvement with their healthcare and is considered a crucial component in the improvement of the health system [4]. Engagement may be indicated by, among other things, increased compliance with treatment regimens and increased satisfaction with care [5] – characteristics which are also associated with improved health outcomes [6]. Patient engagement is especially pertinent to telehealth, since programs require patients to independently interact with their device.

This work extends our previous work [7]. Our goal was to develop predictive models which can prioritize patients for enrollment in the Health Buddy System (HBS), a telehealth program developed by Robert Bosch Healthcare. First, we were interested in predicting the individual cost savings for patients who were enrolled in HBS. This would provide the most direct insight into expected impact from the perspective of a payer. Second, we built models to predict which patients will be highly engaged, which in itself is associated with better health outcomes. Engagement was measured using patient compliance and satisfaction with the telehealth system.

Evaluating cost savings from a non-randomized healthcare program requires consideration of *selection bias*, wherein program participants are different from non-participants with respect to some features, measureable or immeasurable, due to the nature of program enrollment [8]. In healthcare economics, cost savings are often evaluated on a population level using a difference-in-differences approach, propensity scores, instrumental variables, or a combination of these methods [8], which specifically address analytical challenges such as selection bias. However, most such studies lack a more detailed analysis of the cost-savings at an individual level. At the individual level, prediction is especially challenging due to the distribution of medical costs. Costs are skewed right, with high costs driven largely by inpatient hospitalizations [9]. Hospitalizations are relatively rare and can be an important indicator of disease acuity; however, they may also occur as a result of non condition related events such as physical injuries, adding noise to the model [9]. Most *first principle* based approaches, typically adopted in healthcare research, are not well-tailored to handle such complex dependencies [10]. This motivates the need to adopt a data-driven approach towards estimating a generalizable cost-savings model for patients at an individual-level. Many machine learning methods have been applied previously to healthcare data, including the prediction of healthcare costs [11]. However, we found no literature pertaining to prediction of individual-level cost savings or patient engagement for a health intervention. In this paper we first estimate a data-driven cost-savings model for patients at an individual level, and use this model to identify the patients likely to save through the HBS telehealth program. We then estimate a patient engagement model as a supplement to the cost-savings approach. The rest of the paper is organized as follows. Section II describes the cost-savings classifier and the results obtained through applying the model. Section III provides the patient engagement modeling approach and results. Finally the conclusions are presented in section IV.

II. Binary Cost Savings Model

A. Health Buddy Demonstration Study

All data came from the Care Management for High Cost Beneficiaries (CMHCB) demonstration study conducted by the Centers for Medicare and Medicaid Services (CMS) in conjunction with Robert Bosch Healthcare [12, 13]. The study consisted of 11,570 Medicare patients who were offered the Health Buddy to manage their chronic conditions over a period of three years (see http://innovation.cms.gov/Files/reports/CMHCB-HealthBuddyMontefiore.pdf). For this work we utilized only two years; *baseline year* as the year immediately preceding the start of the intervention, and *demonstration year* as the first year of intervention. Further, we removed patients with insufficient data, ESRD, and those whose death falls within six months from the end of the study. The current analysis includes 2383 cases who used the Health Buddy device at least once in the study period and 5092 controls who did not receive a Health Buddy device (see Table I).

Patients came from two major cohorts. The West cohort made up 23% of the total patient sample and included Bend Memorial Clinic in Bend, Oregon and Wenatchee Valley Medical Center in Wenatchee, Washington. Patients in the West cohort were relatively homogeneous and lived in a suburban/rural setting. In contrast, the East cohort contained patients from Montefiore Medical Center in the Bronx, NY, and was demographically diverse in addition to being an urban setting. For the binary cost savings classifier, cohorts were combined in order to maximize power.

B. Data Preprocessing

Data came from two sources:

- *Telehealth Utilization Data* was used to determine program participation and compliance.
- *Administrative Claims Data* contained information on demographics, medical expenditures, and claim counts for baseline and demonstration year.

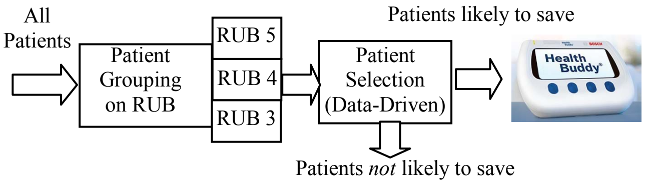


Fig.1. Schematic representation of the cost-savings' modeling workflow.

Claims also included medical diagnosis and procedure codes. We used the Johns Hopkins ACG® system [14] to process claims into meaningful features for medical conditions and healthcare resource utilization (HRU). Of particular importance was the patient grouping information based on morbidity and related health resource utilization patterns, called resource utilization bands (RUBs), which ranged from healthy users with low HRU (RUB 0) to users with complex conditions with very high HRU (RUB 5).

We define *cost savings* as the difference between actual expenditures incurred by patients on HB program and expected expenditures of the same patients had they not been on the HB program. Actual expenditures of telehealth patients can be calculated directly from the claims data; however, expected expenditures need to be estimated from patients who had not participated in the telehealth program (control group), and who are similar to cases with respect to selected features (measured and not measured). As discussed before, the computation of cost savings presented two main challenges: (1) selection bias between cases and controls, and (2) high variability of cost among similar patients caused mostly by variability in hospitalizations and emergency room admissions.

We addressed these issues by comparing the costs of each HBS patient with the average cost within the group of controls who share the same RUB. We then built separate cost savings models for each RUB, viz., RUB 5, RUB 4 and RUB 3 (shown in Fig. 1). The assumption in this approach is that the selection bias is addressed by comparing patients with similar HRU needs and disease conditions. Furthermore, by aggregating the control patients we reduce the variability in the computation of the cost savings and in turn, the variability in model estimation.

Equation (1) provides the difference-in-differences computation we used to estimate cost savings. For example, in the case of a Health Buddy patient in RUB 5 we obtain the cost saving for the i^th patient as:

$$Cost\_Saving_i = \begin{cases} \text{yes} \ ; \ \text{if } (\bar{y}_2^{Control} - y_{i,2}^{HB}) - (\bar{y}_1^{Control} - y_{i,1}^{HB}) \geq 0 \\ \text{no} \ ; \ \text{else} \end{cases} \qquad (1)$$

where, $\bar{y}_2^{Control}$, $\bar{y}_1^{Control}$ is the mean cost of control patients in RUB5 during the demonstration and the baseline year, respectively, and $y_{i,2}^{HB}$, $y_{i,1}^{HB}$ is the cost incurred by the i^th

TABLE I. PATIENT DEMOGRAPHICS

	Controls		Cases				Controls		Cases		
Variable	Mean/ Freq	SD/%	Mean/ Freq	SD/%	P*	Variable	Mean/ Freq	SD/ %	Mean/ Freq	SD/ %	P*
N	5092		2383			***Morbidity***					
Demographics						AHRQ [15]	4.07	2.24	4.04	2.13	0.980
Cohort					0.000	Charlson [16]	2.38	1.55	2.39	1.49	0.510
East	3901	77%	1200	50%		Elixhauser [17]	4.33	2.35	4.30	2.24	0.845
West	1191	23%	1183	50%		Chronic condition count	5.87	3.21	6.87	3.26	0.000
Age	77.95	9.30	75.27	9.24	0.000	Condition categories					
Sex					0.003	Congestive heart failure	1039	20%	548	23%	0.011
Male	2080	41%	1061	45%		Diabetes	2167	43%	1198	50%	0.000
Female	3012	59%	1322	55%		COPD	911	18%	579	24%	0.000
Race					0.223	***Resource Utilization***					
Unknown	19	0%	7	0%		Inpatient hosp. counts	0.44	0.83	0.71	1.15	0.000
White	3649	72%	1672	70%		Emergency dept. visit count	0.36	0.89	0.62	1.30	0.000
Black	850	17%	436	18%		Outpatient visit count	25.5	19.2	37.8	26.0	0.000
Other	120	2%	46	2%		Nursing service	354	7%	102	4%	0.000
Asian	50	1%	15	1%		Major procedure	363	7%	399	17%	0.000
Hispanic	381	7%	197	8%		Cancer treatment	191	4%	97	4%	0.504
NA Native	23	0%	10	0%		Frailty flag	1423	28%	692	29%	0.328
Age Group					0.002	Resource utilization band					0.000
Over 65	4828	95%	2216	93%		Under 3	68	2%	10	0%	
Under 65	264	5%	167	7%		RUB 3	1324	26%	410	17%	
Costs and Claims						RUB 4	1816	36%	810	34%	
Cost in Y1	15369	19074	17243	19847	0.000	RUB 5	1884	37%	1153	48%	
Cost in Y0	15150	18779	17521	19046	0.000	***Predicted Resource Utilization***					
						Probability high total cost	0.04	0.04	0.05	0.05	0.000
						Probability inpatient hosp.	0.26	0.15	0.31	0.17	0.000

*Continuous variables tested with Wilcoxon rank-sum test; categorical variables tested with Pearson chi-squared test.

HBS patient in RUB 5 during the demonstration and the baseline year, respectively.

We build binary classifiers of "+1" (save) and "-1" (loss) classes (i.e., the cost-savings model) separately for each RUB and identify the patients who are likely to save (see Fig.1). After discussions with application domain experts, we selected 40 variables (listed in Table II) for modeling. These variables capture the demographics, clinical, and historical claims information for each patient. Next, the data was uniformly scaled in the range of [0, 1]. The final dataset used for modeling is the result of the selection and preprocessing steps detailed above. Table III contains final class sizes within each RUB.

C. Modeling

To build the cost savings classifier we tried several approaches including Decision Trees, L2- regularized Hinge Loss Support Vector Machine (SVM), L1-regularized L2-Loss SVM, L1-regularized Logistic Regression, and L2-regularized Logistic Regression (see [10] for more details). Since all of the methods provided similar classification accuracy, we settled on the L2-regularized Logistic Regression because it provides two main advantages:

- L2-regularization controls the model complexity and results in more generalizable models (i.e. models which can avoid over fitting). In addition, solving the L2-regularized Logistic Regression is more tractable in comparison to the other L1 based approaches.
- The logit loss provides a probabilistic output, which results in a more interpretable model in comparison to the SVM based approaches.

The L2-regularized Logistic Regression formulation is

TABLE II. FEATURE WEIGHTS FOR DIFFERENT RUBS

FEATURES	RUB 5	RUB 4	RUB 3
Inpatient payments	5.53	5.95	1.48
Race = Asian	-4.93	-0.49	-7.50
Carrier payment	4.58	8.50	0.34
Outpatient payment	2.79	4.08	1.85
Race = N. Am. Native	-2.13	-1.16	1.37
HHA payment	1.98	-0.05	-19.0
DME count	1.95	-0.46	0.17
DME payments	-1.82	-0.01	2.40
Inpatient claim count	1.65	-2.05	1.80
Baseline year cost	1.53	2.07	0.13
Total no. of claims	-1.34	-2.21	-3.11
Race = Other	-1.16	0.00	3.16
CHF claim count	-1.03	1.15	-1.60
Charlson comorbidity	-0.96	-0.35	-2.65
COPD claim count	-0.96	2.00	4.43
Difference from RUB mean for Y1	-0.72	0.14	-0.16
Race = Black	-0.56	0.00	2.75
SNF claim count	-0.53	4.12	14.7
Race = Hispanic	-0.42	-0.56	0.95
East Phase I	-0.42	-0.11	-0.35
Race = White	-0.4	0.35	2.41
East Phase II	-0.39	0.44	0.83
CHF Diagnosis	0.38	-0.54	0.02
Age over 65	-0.36	0.16	-0.13
Elixhauser comorbidity	-0.33	-0.34	0.88
Diabetes claim count	-0.32	-0.33	0.74
Age under 65 (disabled)	-0.21	0.39	0.17
Female	0.17	-0.09	-0.02
Age	-0.12	-0.13	-0.85
West Phase II	-0.08	-0.10	-0.19
Outpatient claim count	0.07	-0.18	1.63
AHRQ cormorbidity	-0.07	-0.13	-0.44
Diabetes diagnosis	-0.05	-0.12	-0.21
West Phase I	-0.04	0.33	-0.29
SNF payments	-0.04	-6.07	0.00
Carrier claim count	0.03	0.33	1.17
COPD diagnosis	0.03	-0.31	-0.62
HHA claim count	0	0.46	10.3

TABLE III. PATIENT COUNTS BY CLASS AND RUB

	RUB 5	RUB 4	RUB 3
Number of patients	1155	812	417
Number of patients with cost savings (class '+1)	476 (41%)	497 (61%)	296 (71%)
Number of patients without cost savings (class '-1')	679 (59%)	315 (39%)	121 (29%)

TABLE IV. PERFORMANCE OF THE ESTIMATED MODEL OVER (5,5) DOUBLE RESAMPLING

Performance Metric	Test	Training
RUB 5		
Wt. Error Rate (%)	20.87 (3.04)	18.90 (0.78)
Proportion Selected (%)	51.91 (2.68)	51.56 (0.69)
Cost Saving (after selection)	$10,200 ($2,037)	$10,359 ($748)
Cost Saving (before selection)	-$1,886 ($851)	-$2,324 ($211)
RUB 4		
Wt. Error Rate (%)	39.92 (4.59)	34.06 (1.23)
Proportion Selected (%)	48.83 (4.44)	50.35 (1.99)
Cost Saving (after selection)	$5,866 ($2,628)	$6732 ($923)
Cost Saving (before selection)	$225 ($1,279)	-$35 ($318)
RUB 3		
Wt. Error Rate (%)	40.43 (4.78)	31.34 (1.81)
Proportion Selected (%)	59.16 (3.52)	58.95 (1.95)
Cost Saving (after selection)	$2,914 ($1,174)	$4,532 ($434)
Cost Saving (before selection)	-$437 ($1,794)	$8 ($446)

defined next [10]. Given input training data $(\mathbf{x}_i, y_i)_{i=1}^{N}$ with $\mathbf{x} \in \Re^D$ and $y \in \{-1,+1\}$, solve:

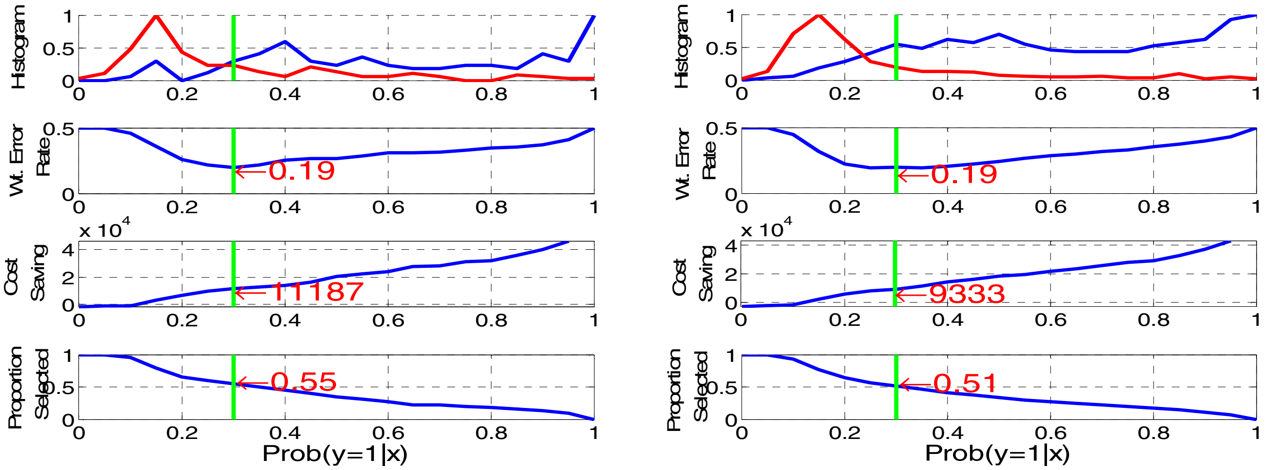
$$\min_{\mathbf{w},b} \quad \underbrace{\frac{1}{2}\|\mathbf{w}\|_2^2}_{\text{L2-Regularizer}} \quad + \quad \underbrace{\frac{C}{N}\sum_{i=1}^{N}\log(1+e^{-y_i(\mathbf{w}^T\mathbf{x}_i+\mathrm{b})})}_{\text{logit loss}} \qquad (2)$$

where N represents the total number of training samples, D is the dimension of input samples, $C \geq 0$ and θ are user-defined parameters typically selected on the basis of application domain knowledge. In this paper the $C \geq 0$ and θ parameters have been selected through model selection. The output is a probabilistic model with the final decision rule given as:

$$D(\mathbf{x}) = \begin{cases} +1, \text{ if } P(y=+1\,|\,\mathbf{x}) = \dfrac{1}{(1+e^{-(\mathbf{w}^T\mathbf{x}_i+b)})} \geq \theta \\ -1, \text{ else} \end{cases} \qquad (3)$$

D. Results

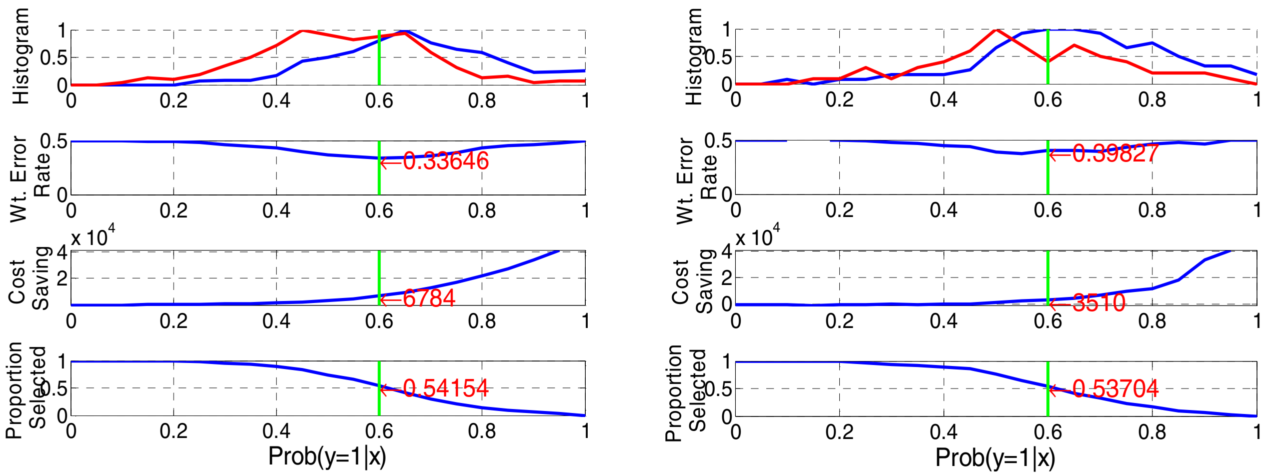
We provide the experimental results for (5,5) double resampling. Double resampling is a machine learning technique typically used to evaluate the predictive power of an algorithm. This approach involves a two-level partitioning of the data. The inner partition is used to perform model selection (i.e. selection of the optimal model parameters), in this case $C \geq 0$ and θ. The outer partition is used to test the predictive power of the selected optimal model. The resampling is done multiple times (while maintaining the same prior probabilities for each partition), and the average weighted error rate over the several partitionings is reported in Table IV together with the



(a)

(b)

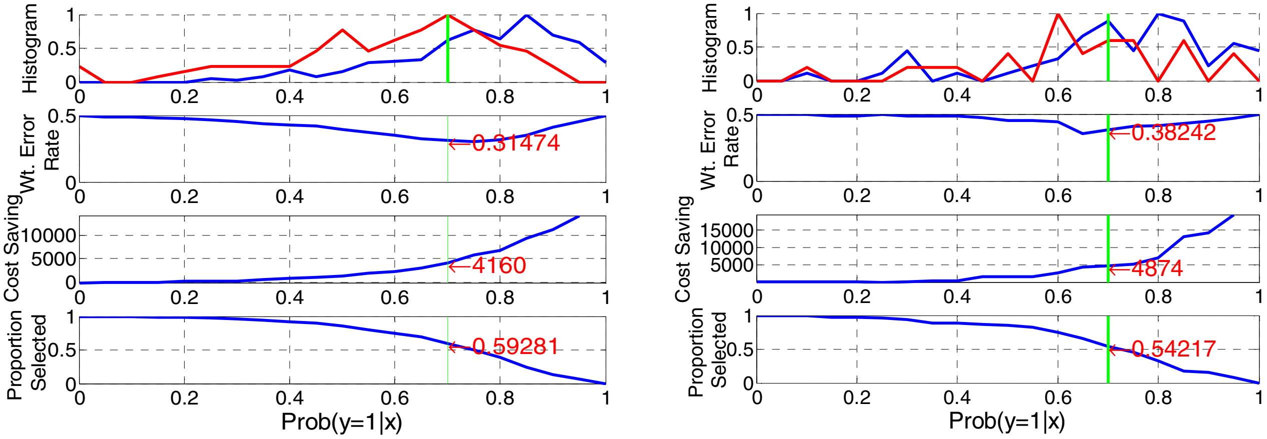
Fig.2. Model Performance for RUB 5 with C=10, θ = 0.3 for one random partition of the 5-fold split. (a) training data. (b) test data.



(a)

(b)

Fig.3. Model Performance for RUB 4 with C=100, θ = 0.6 for one random partition of the 5-fold split. (a) training data. (b) test data.



(a)

(b)

Fig.4. Model Performance for RUB 3 with C=1000, θ = 0.7 for one random partition of the 5-fold split. (a) training data. (b) test data.

standard deviations (in parenthesis). The weighted error $C_{fp}P_{fp} + C_{fn}P_{fn}$, takes into consideration the imbalance nature of the dataset. Here, P_{fp} and P_{fn} denote the probability (error rate) of false-positives and false-negatives errors (respectively); and C_{fp}, C_{fn} are the costs of misclassifications associated with the false-positives and false-negatives errors. For empirical comparisons this weighted error is normalized by its maximum possible value ($C_{fp} + C_{fn}$) as shown next,

$$\text{Normalized}\, C_{fp}P_{fp} + C_{fn}P_{fn} = \frac{r(n_{fp}/n^-) + (n_{fn}/n^+)}{r(n^-/n^-) + (n^+/n^+)}$$

$$= \frac{r(n_{fp}/n^-) + (n_{fn}/n^+)}{r+1}\text{; with } r = C_{fp}/C_{fn}\,.$$

Here, n_{fp}, n_{fn} denotes the number of false positives and false negative samples, and n^+, n^- denotes the number of positive and negative samples. Such a normalization limits the range of the error to [0, 1], which is the same range as standard binary classification problems (with equal costs and prior class probabilities). In the rest of the paper we refer to this normalized weighted error rate as simply the weighted error rate.

Figs. 2-4 provides the performance of the models for the typically selected parameters over one outer-partitioning of the resampling technique. Based on the results from Table IV (see Figs 2-5), with 50% of the patients selected we have,

- *RUB 5*: the estimated model (shown in Fig. 2 and Table II), can save ~ \$10,000 (approximately) in comparison to the current loss of ~ - \$2000 (approximately).
- *RUB 4*: the estimated model (shown in Fig. 3 and Table II), can save ~ \$6,000 (approximately) in comparison to the current saving of ~ \$100 (approximately).
- *RUB 3*: the estimated model (shown in Fig. 4 and Table II), can save ~ \$3000 - \$5000 (approximately) in comparison to the current loss of ~ - \$500 (approximately).

Further, the estimated models shows similar performance (weighted error, cost savings etc) for both training/test data, indicating a good model fit.

Finally, based on our analyses we can infer that such a data-driven filtering approach can prove useful in selecting at least 50% of the patients with reasonably high cost-savings. Finally, feature selection could shed more light on the effective variables to build a better predictive model. For example, in the case of RUB 5, we see that features such as inpatient payment, race, and carrier payment are important for selecting the patients who are likely to save, while features like number of home health agency claims, number of hospice claims, and hospice payments are the least predictive for building such a cost savings model (Table II). As a word of caution, it is not recommended to associate a causal behavior between cost-savings and the above features based on the estimated model. For such steps appropriate experiments proving causal relationship need to be set up.

TABLE V. ITEMS FROM THE 90-DAY SATISFACTION SURVEY

S1. Satisfaction with Health Buddy program Overall how satisfied are you with the Health Buddy? 5 = Very satisfied, 4 = Satisfied, 3 = Somewhat satisfied, 2 = Not very satisfied, 1 = Not at all satisfied
S2. Improvement in disease knowledge Since I started answering the educational questions on the Health Buddy, my understanding of my medical condition is: 5 = Much better, 4 = Somewhat better, 3 = Neutral, 2 = Somewhat worse, 1 = Much worse
S3. Improvement in self-management Since I started answering the educational questions on the Health Buddy, I am able to manage my medical condition: 5 = Much better, 4 = Somewhat better, 3 = Neutral, 2 = Somewhat worse, 1 = Much worse
S4. Willingness to recommend to a friend How willing are you to recommend the Health Buddy to others? 5 = Very willing, 4 = Somewhat willing, 3 = Neutral, 2 = Somewhat unwilling, 1 = Very unwilling

III. Patient Engagement Model

The goal of the patient engagement model was to discover the relationships between predictor variables such as claim costs, demographics or diagnostic information and patient engagement with the Health Buddy device. We considered patient compliance and patient satisfaction as proxies for engagement and built separate models for each outcome. Such a model can supplement the "cost-savings" based patient selection model, by integrating the additional information of the individual patient's engagement.

A. Data Source.

We used the same dataset from the cost savings model (section IIa and IIb), restricted to patients who had been enrolled in the Health Buddy for at least one year. The resulting dataset included 1483 patients and 30 features for consideration. In addition we considered *patient satisfaction surveys* (shown in Table V), administered during the 90th telehealth session, as both a predictor of compliance and as an outcome itself. Finally, in addition to creating models for the combined cohorts, we also looked separately at the East and West cohorts due to the demographic differences.

B. Calculation of Compliance.

Patient compliance measures how regularly a patient used the Health Buddy device over the course of an year, and is calculated as the ratio of the number of times a patient used his/her device over 365 (the number of days in the year). A ratio of 1 indicates a highly compliant

TABLE VI. SIGNIFICANT FEATURES FOR COMPLIANCE MODEL (WITH 75% THRESHOLD)

Feature	OR	2.50%	97.50%	P-Value
All Cohorts ***(n = 1483, 48% compliant)***				
Phase				
East Phase 1	1.00 (Ref)	-	-	-
East Phase 2	2.55	1.75	3.74	0.00
West Phase 1	1.74	1.20	2.53	0.00
West Phase 2	1.17	0.77	1.76	0.47
Sex = Female	0.64	0.50	0.81	0.00
Claim Count	6.89	2.19	22.09	0.00
COPD	1.57	1.19	2.09	0.00
Hypertension	0.71	0.53	0.94	0.02
East Cohort ***(n = 610, 43% compliant)***				
East Phase 2	2.37	1.56	3.65	0.00
Sex = Female	0.44	0.29	0.67	0.00
Claim Count	5.69	1.04	33.21	0.05
Last 3 months cost	0.10	0.01	0.84	0.04
West Cohort ***(n = 873, 53% compliant)***				
Last 3 months cost	5.50	1.07	30.43	0.05
Last 6 months cost	0.05	0.00	0.49	0.01
COPD	2.00	1.38	2.92	0.00
Hypertension	0.69	0.48	0.98	0.04
Lower Back Pain	0.68	0.50	0.93	0.02

TABLE VII. SIGNIFICANT FEATURES FOR SATISFACTION MODEL

Feature	OR	2.50%	97.50%	P-Value
All Cohorts ***(n=1483, 24% highly satisfied*)***				
Chronic renal failure	1.64	1.14	2.34	0.01
Diabetes	1.34	1.02	1.77	0.04
Chronic condition count	0.17	0.03	0.91	0.04
East Cohort ***(n=610, 25% highly satisfied*)***				
Chronic condition count	0.03	0.00	0.34	0.01
Age 75-79	0.46	0.22	0.95	0.04
West Cohort ***(n=873, 23% highly satisfied*)***				
Chronic renal failure	1.81	1.01	3.19	0.04

TABLE VIII. AUC FOR ENGAGEMENT MODELS

Model	All cohorts	East	West
Compliance Models			
Compliance at 75%	0.620	0.610	0.607
(with Total Satisfaction)	0.638	0.616	0.641
(with S1)	0.642	0.640	0.633
Compliance at 40%	0.592	0.609	0.568
(with Total Satisfaction)	0.604	0.619	0.598
(with S1)	0.600	0.615	0.578
Satisfaction Models			
Total Satisfaction - All 5's	0.566	0.571	0.559
Total Satisfaction - 4's & 5's	0.535	0.573	0.551
S1 only, for 4 & 5	0.560	0.631	0.584

individual while a ratio of 0 indicates a non-compliant individual. Since our interest is in compliance vs. non-compliance rather than the degree of compliance, a threshold of 0.75 was placed on this ratio to convert it into a binary (compliant/non-compliant) result. That is, patients with a compliance score of 0.75 and greater were marked as complaint, while those with a score less than 0.75 were marked as non-compliant. This threshold was selected after discussions with domain experts. Further, we also tried a compliance threshold of 0.40, based on a previous analysis in which this was the median compliance level [6]. However, in our analysis, we found that only 18% percent of the population fell below the 0.40 threshold. This is likely due to the range of analysis; which is the first year in this case.

C. *Definition of Satisfaction.*

Patient satisfaction, measures how satisfied a patient is with his/her device and is calculated based on the numerical responses given to a set of five satisfaction questions that patients answer in the 90th Health Buddy session. The survey responses for each question ranged from 1 – 5 with 1 being extremely satisfied and 5 being extremely dissatisfied (also see Table V). We constructed two sets of satisfaction scores from these responses. The total satisfaction, SAT, is defined as:

$$SAT = 24 - (S1 + S2 + S3 + S4) \tag{4}$$

We also experimented with using only the response to first question since it measured the patient's general satisfaction with the HB device and program. That score, referred to as SAT1, is defined as:

$$SAT1 = 6 - S1. \tag{5}$$

By subtracting the response numbers from 24 (or 6 in the case of SAT1) we are able to convert the numbers to a more understandable scale with the higher numbers indicating higher satisfaction with the HB device.

D. *Development of Patient Engagement Model.*

All patient engagement models used logistic regression. Model covariates included socio-demographic information, claims data, output from the ACG system, and a limited selection of questions from the Health Buddy content. Model performance was assessed using the area under the curve (AUC) with cross-validation. The significant features for the estimated models for both the compliance (with 75% threshold) and the satisfaction models are provided in Tables VI and VII respectively.

E. *Results*

The final performance of the estimated models is provided in Table VIII. As seen from Table VIII, there was greater predictive performance for the model with a 75% compliance threshold (AUC = 0.620) than the model with 40% compliance threshold (AUC = 0.592).

Further, high total satisfaction or overall satisfaction (S1) were predictive of compliance and improved the performance of the model. Further, as seen from Table VI, several features were associated with high patient compliance including cohort, sex, COPD, hypertension, and number of claims. However, no features show consistent, significant effects in both the East and West cohorts. This is also consistent for the estimated model with threshold of 40% (and hence been omitted from the paper).

Finally, for the satisfaction models, chronic renal failure, diabetes, and a lower chronic condition count were associated with higher satisfaction (see Table VII). Once again, there were no overlapping significant predictors between the East and West cohorts. The overall satisfaction question had predictive performance comparable, or even exceeding, that of the total satisfaction scores (as seen from Table VIII).

IV. Summary

The problem of estimating individual-level cost savings for health interventions is difficult due to the identification of an appropriate comparison group and adjustment for selection bias. Furthermore, the variabiliy of medical costs, driven largely by hospitalizations, added additional challenges. We computed ground truth by comparing Health Buddy patients with an aggregated group of controls with similar resource utilization and estimated separate logistic regression models for each RUB. Filtering the patients based on the predictive model resulted in high cost savings in comparison to the current low savings/loss incurred by the Health Buddy system. Further, our models used linear parameterization, which is easy to interpret and can be tuned to control the interplay between selected patient population size and targeted cost savings per customer's expectations.

Several features were identified as predictive of patient compliance and satisfaction, but results were not consistent across cohorts. The 75% compliance model (AUC 0.62-0.64) had better performance than the satisfaction model (AUC 0.54-0.57). However, these levels of accuracy are generally not sufficient for implementation as part of a patient identification tool.

Results of the patient engagement models do provide directions for future work. All models were sensitive to patient cohort, which is not particularly useful within the context of patient identification for a defined population, but does support the notion that the health system and/or geographic region of a patient is important to their success with telehealth. In future studies, having detailed notes on the involvement of nurses and their interaction with patients may help bring greater insights into compliance. Furthermore, the four cohorts were different with respect to not only medical conditions, but also demographic, socioeconomic, and cultural characteristics, which may also be implicated in the level of patient engagement. Finally, additional validated measures which specifically target patient behavior, such as the Patient Activation Measure [18] might provide greater insight to engagement than medical and demographic features alone.

Finally, it is important to note that associating causal relationship through the estimated predictive models should be done with caution [19]. This paper does not provide a causal-effect analysis relating the patient's demographics, clinical and historical claims information to the cost-savings. However, when this approach is complemented with domain knowledge, insights regarding the features necessary/unnecessary to prediction of cost savings can still be derived. As an example, for the cost savings model, inpatient payments, which exemplify regression to the mean, could create an impression of higher cost-savings. Future research could exclude such features. Moreover, the current model does not incorporate the cost-values into the loss function of the model. Incorporating the cost-values could further help to identify patients with high vs. low cost savings. This could yield a better patient identification model which additionally weighs the patients in terms of the amount they save.

Future efforts may benefit from estimating cost savings for groups of patients rather than individuals, similar to the actuarial cell approach used by clinical groupers to predict resource use. This would further mitigate the negative effect of cost variation on individual patients' estimated cost savings. In addition, using a comparison group that was not offered the Health Buddy would further reduce the selection bias.

Acknowledgment

The authors would like to thank Dr. Michael Finch of the University of Minnesota, Donald Hughes of Encompass Medical Management, Inc., and Bill Broderick of Robert Bosch Healthcare Systems, Inc. for useful discussions.

References

[1] L. C. Baker, S. J. Johnson, D. Macaulay and H. Birnbaum, "Integrated Telehealth and Care Management Program for Medicare Beneficiaries with Chronic Disease Linked to Savings," *Health Affairs,* vol. 30, no. 9, pp. 1689-1697, 2011.

[2] S. Weir, G. Aweh and R. E. Clark, "Case Selection for a Medicaid Chronic Care Management Program," *Health Care Financing Review,* vol. 30, no. 1, pp. 61-74, 2008.

[3] J. A. Fleishman and J. W. Cohen, "Using Information on Clinical Conditions to Predict High-Cost Patients," *Health Services Research,* vol. 45, no. 2, pp. 532-552, 2010.

[4] K. L. Carman, P. Dardess, M. Maurer, S. Sofaer, K. Adams, C. Bechtel and J. Sweeny, "Patient and Family Engagement: A Framework for Understanding The Elements And Developing Interventions And Policies," *Health Affairs,* vol. 32, no. 2, pp. 223-231, 2013.

[5] A. Coulter, "Patient Engagement--What Works?," *Journal of Ambulatory Care Management,* vol. 35, no. 2, pp. 80-89, 2012.

[6] W. C. Broderick, K. Gilberg and D. Macaulay, "The Effects of Compliance to Daily Telehealth Sessions on Hospitalisation Rates in the Chronically Ill Elderly," in *International Journal of Integrated Care*, London, UK, 2013.

[7] M. Ganser, S. Dhar, U. Kurup, C. Cunha and A. Gacic, "Patient Identification for Telehealth Programs," in *ICMLA*, Miami, 2015.

[8] S. R. Khandker, G. B. Koolwal and H. A. Samad, Handbook on Impact Evaluation, Washington, D.C.: The World Bank, 2010.

[9] P. Diehr, D. Yanez, A. Ash, M. Hornbrook and D. Y. Lin, "Methods For Analyzing Health Care Utilization And Costs," *Annual Review of Public Health,* vol. 20, pp. 125-44, 1999.

[10] V. Cherkassky, Predictive Learning., VCtextbook , 2013.

[11] D. Bertsimas, M. V. Bjarnadottir, M. A. Kane, J. C. Kryder, R. Pandey, S. Vempala and G. Wang, "Algorithmic Prediction of Health-Care Costs," *Operations Research,* vol. 56, no. 6, pp. 1382-1392, 2008.

[12] C. Urato, N. McCall, J. Cromwell, N. Lenfestey, K. Smith and D. Raeder, "Evaluation of the Extended Medicare Care Management for High Cost Beneficiaries (CMHCB) Demonstration: Health Buddy Program at Montefiore," RTI International, Research Triangle Park, NC, 2013.

[13] C. Urato, N. McCall, J. Cromwell, N. Lenfestey, K. Smith and D. Raeder, "Evaluation of the Extended Medicare Care Management for High Cost Beneficiaries (CMHCB) Demonstration: Health Buddy West Program," RTI International, Research Triangle Park, NC, 2013.

[14] J. P. Weiner and C. Abrams, Eds., The Johns Hopkins ACG System Technical Reference Guide, Version 10.0, Baltimore, Maryland: Johns Hopkins Bloomberg School of Public Health, 2011.

[15] A. Elixhauser, C. Steiner and D. Kruzikas, "Comorbidity Software Documentation, HCUP Methods Series Report # 2004-1.," U.S. Agency for Healthcare Research and Quality., 6 February 2004. [Online]. Available: http://www.hcup-us.ahrq.gov/toolssoftware/comorbidity.comorbidity.jsp. [Accessed 26 August 2015].

[16] R. A. Deyo, D. C. Cherkin and M. A. Ciol, "Adapting a Clinical Comorbidity Index For Use With ICD-9-CM Administrative Databases," *Journal of Clinical Epidemiology,* vol. 45, no. 6, pp. 613-9, 1992.

[17] A. Elixhauser, C. Steiner, D. R. Harris and R. M. Coffey, "Comorbidity Measure For Use With Administrative Data," *Medical Care,* vol. 36, pp. 8-27, 1998.

[18] J. H. Hibbard, J. Stockard, E. R. Mahoney and M. Tusler, "Development of the Patient Activation Measure (PAM): Conceptualizing and Measuring Activation in Patients and Consumers," *Health Services Research,* vol. 39, no. 4, pp. 1005-1026, 2004.

[19] V. Cherkassky and S. Dhar, "Interpretation of Black-Box Predictive Models," in *Measures of Complexity: Festschrift for Alexey Chervonenkis*, 2016.