# Uplift Models: Can They Be Used to Identify and Rank Heart Failure Patients Expected to Benefit From a Clinical Telehealth Program?

Atilla Asar
STUDENT NUMBER: 2031501

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN COGNITIVE SCIENCE & ARTIFICIAL INTELLIGENCE
DEPARTMENT OF COGNITIVE SCIENCE & ARTIFICIAL INTELLIGENCE
SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
TILBURG UNIVERSITY

Thesis committee:

dr. Marijn van Wingerden
prof. dr. Steffen Pauws

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science & Artificial Intelligence
Tilburg, The Netherlands
June 2019

**Preface**

Firstly, I would like to thank prof. dr. Steffen Pauws for giving me the opportunity to be involved in this research and for his continuous support. I would also like to thank dr. Marijn van Wingerden for his valuable feedback. To dr. Tobias Wirth, I am very grateful for the insightful discussions. Finally, I want to thank Philips for giving me the opportunity to work with the dataset and the support.

**Table of Contents**

# Uplift Models: Can They Be Used to Identify and Rank Heart Failure Patients Expected to Benefit From a Clinical Telehealth Program?

Atilla Asar

*It is of importance to identify people with chronic or life-threatening diseases that can be recruited for particular treatments where they will receive optimal benefits and health gains in the field of health care. In this study, we assessed treatment response uplift modeling to motivate the take-up of uplift modeling in clinical settings. We presented the applications of four different uplift modeling techniques on a real-world dataset from a randomized clinical trial testing an intervention in a heart failure patient cohort, TEN-HMS study (The Trans-European Network-Home-Care Management System) and a synthetic dataset comprising an intervention and control in synthesized heart failure patient cohort. We demonstrated and evaluated the implementation of two-model, dummy treatment, pessimistic, and generalized Lai's approaches. We made performance and stability experiments which show that generalized Lai's approach (glai) is the model of choice with consistently higher performance and stability, in comparison to the other three methods. The stability experiments were performed with five different random samples with half of the observations of our generated dataset. The implementation of the glai approach demonstrated that recruitment of the patients according to their NYHA (New York Heart Association), BMI (Body Mass Index), and a number of prior hospitalizations can be used together. We demonstrated that uplift modeling could be used to identify a subgroup of patients with heart failure who will benefit most from an HTM intervention based on its clinical characteristics.*

## 1. Introduction

### 1.1. Problem Statement

Randomized clinical trials (RCTs) are the gold standard for evidence-based medicine to test the effect of a treatment on a specified outcome. The main result of an RCT gives the average effect of treatment across the tested population (Deaton 2018). Even if RCTs provide good evidence of treatment effects on a population, this can be misleading on an individual level. In some significantly positive trials, some of the treated patients, both within and outside the clinical trials, may receive little or no benefit, or even harm from a treatment (Kent 2007). Such nonrandom and explainable variations in treatment effects, called heterogeneity of treatment effects, raises an important question: "are there subgroups of patients in a trial who are more likely or unlikely to derive benefit from a treatment, and can we stratify the ones who will receive the optimal benefit?"

A common strategy used in medical sciences to investigate the heterogeneity of treatment effects is subgroup analysis. Subgroup analysis is the evaluation of treatment effects across a

number of subgroups of patients using a particular variable, one variable at a time, or a combination of variables (Meinert 2009)(pg 288-290). There are many limitations for the subgroup analyses (Austin 2006). In reality, there can be myriad subgroup analyses possible in any given RCT, thus researchers have to use their expertise in the field to choose the variables, which introduces a bias in the feature selection. Furthermore, analyzing each of these variables "one variable at a time" (e.g., male vs female, old vs young) will risk spurious false-positive subgroup results. As a result, subgroup analyses mostly ignore the joint influence of covariates and may fail to detect clinically significant multiple covariates causing the variations in treatment effects (Kornegay 2013).

The traditional machine learning methods can be applied on RCTs; however, the essential problem with them is that they do not identify the subgroups of patients that would benefit most from an intervention. The traditional classification methods only predict the probability of responding rather than the increase in the probability of responding based on intervention. In other words, conventional classification methods can only model what happens after the intervention, which individuals will experience a particular outcome, but cannot model what happens due to the intervention, which patient or subgroup of patients will benefit from a specific treatment. The main reason for this shortcoming is that these models do not take into account what would have happened if the intervention were not implemented. On the other hand, during a standard randomized clinical trial, a random group of patients is subjected to treatment and another random group to an alternative treatment. Random distribution guarantees that there will be no systematic differences in factors, both known and unknown, that may affect the outcome (Sibbald 1998). The patients assigned to the treatment to be tested is called the treatment group, whereas the other counterpart is called the control group. In contrast to traditional classification methods, uplift modeling allows for the presence of control groups and can predict which individual patients will benefit from a specific treatment based on their specific characteristics. This is why we chose uplift modeling in this project to address which subgroup of patients with heart failure will get the optimal benefit from a remote telemonitoring.

Uplift modeling is a type of predictive machine learning technique, aims of which to detect the incremental gains of an outcome from an intervention on a population (Lo 2015). The predicted results from uplift modeling are unlabeled because, in a randomized experiment, the outcome cannot be observed simultaneously for a single individual (the person either receives a treatment or does not receive it). This is the phenomenon known as the *Fundamental Problem of Causal Inference* (Holland 1986). To overcome this problem, uplift models depends essentially on randomized experiments. A detailed overview of uplift modeling techniques that have been implemented throughout this project is provided in Section 2.1.

### 1.2. Context-of-use

In order to report the viability of the uplift modeling on RCTs, we tested different approaches on two benchmark datasets. The first is the clinical trial data collected during TEN-HMS study (The Trans-Europian Network-Home-Care Management System) which investigated the effect of home telemonitoring treatment effects on patients with heart failure (Cleland 2005). The second

is synthetic dataset retaining the structure of the first one. A detailed overview of the datasets used in this project is provided in Section 3.1.

Heart failure (HF) is a chronic, progressive and complex cardiovascular disorder having high prevalence and incidence worldwide. In the case of untreated or not well-managed conditions, deterioration of HF may require frequent and prolonged hospitalization, which can worsen the prognosis for the disease and the subsequent survival among affected patients (GJ de Vries 2013, Martirosyan 2017, Roth 2015). Globally cardiovascular deaths increased by 41% between 1990 and 2013, whereas age-specific death rates fell by 39% (Roth 2015). Most of the management schemes for Patients with heart failure undergoing standard care consists of close clinical follow-up. Although patients with heart failure undergoing standard care frequently attend the scheduled office visits and follow-ups, unplanned cardiovascular hospital readmissions and mortality rates among these patients are still high (Chioncel 2017). Additionally, these intense face-to-face follow-up strategies are too costly and add extra load on the patients because of their demands on the patient's time and travel needs, which might, as well, limit the number of patients who can participate in such schemes.

Home telemonitoring (HTM) can address above-mentioned issues and might be a benefit for the patients with heart failure and the whole healthcare system with respect to cost efficiencies and clinical effectiveness. HTM is a form of non-invasive, remote patient monitoring strategy which consists of a digital transmission of physiological data e.g. electrocardiogram, blood pressure, weight, pulse oximetry, respiratory rate, and other data (self-care, education, lifestyle modification, and medicine administration) (Stewart 2011). By remotely collecting data on a regular basis, HTM patient management strategy allows health-care providers to monitor patients' symptoms and guide them, using telecommunications as an alternative to or alongside in-person visits. Therefore, caregivers can detect clinical decompensation of patients with heart failure earlier, and take on-time interventions to prevent HF-related mortality cases or further deterioration of the patient condition.

Remote monitoring of patients with heart failure also provides access to specialist care for a much larger number of patients, particularly for those living in remote geographical areas or the frail ones who are housebound, as well as those at high risk of deterioration. A series of recent randomized clinical trials (RCT) indicate that HTM can reduce the proportion of day lost due to unplanned cardiovascular hospital admissions and all-cause mortality risks among patients with heart failure (Koehler 2018, Yun 2017). Most importantly, these studies emphasized that HTM initiated some potentially life-saving hospital admissions, even if it slightly decreases the overall number of HF caused hospital admission days.

Despite the fact that HTM has shown significant improvements in health-related quality of life for patients with heart failure, it is not possible to offer this treatment to every patient (Inglis 2015, Yun 2018). First, as mentioned in the above section, each patient with heart failure might not receive similar beneficial effects from HTM, some patients even might get little to no benefit. Clearly, a one-size-fits-all approach will not work for every patient. Patients are diverse in terms of their demographic and clinical profiles. Second, the cost of recruiting every patient to HTM intervention is not a financially viable option to the standard care (Inglis 2015, McDowell 2015, Williams 2016). Therefore, patients with heart failure need to be ranked due to the aforementioned reasons. If there is an adequate risk stratification among these patients for HTM

recruitment, they may benefit most from this intervention; it can extend their lives in years and can improve their health-related quality of life.

To best of our knowledge, there has been no previous research using uplift modeling to select the optimal treatment for patients with heart failure. As mentioned before, the prediction of the most efficient treatment from HTM and usual care based on patients' characteristics is of special interest. This brings the following research questions to be answered:

- "Can uplift models be used to identify patients with heart failure who benefit most from the home telemonitoring treatment based on clinical benefits?"
- "If so, which uplift model technique will give the most stable and efficient performance in our benchmark datasets?"

Using two benchmark datasets, the TEN-HMS trial and artificially generated synthetic datasets, this thesis applied different uplift modeling techniques to define subgroups of patients with heart failure who will potentially benefit most from the HTM intervention. Moreover, the HTM intervention impact on the selected patients with respect to their hospital admissions and mortality was predicted. The successful application of uplift modeling in this project will show which individual patients with given baseline characteristics have a higher likelihood of receiving a benefit in hospital-free survival from HTM.

The rest of this thesis is organized as follows: Section 2 gives an overview of the related work. Section 3 describes the datasets that we used during this project, the construction of uplift modeling techniques, and their evaluation process. Section 4 presents the experimental results and evaluation of the models. Section 5 discusses the main findings and provides directions for future research. Finally, Section 6 concludes with the list of what has been done during this project and the main findings.

## 2. Related Work

In this section, we first briefly introduce uplift modeling and uplift modeling techniques that have been implemented throughout this project. Subsequently, a comprehensive literature review on uplift modeling is provided.

### 2.1. Uplift Modeling and Uplift Modeling Techniques

Uplift modeling is a type of predictive machine learning technique, aims of which to detect the true differences in the probability of an outcome from intervention in a population of individuals (Lo 2015). The predicted results from uplift modeling are unlabeled because, in a randomized experiment, the outcome cannot be observed simultaneously for a single individual (the person either receives the treatment or does not receive it). This is the phenomenon known as the *Fundamental Problem of Causal Inference* (Holland 1986). To overcome this problem, uplift models depends essentially on randomized experiments. These models can be defined as follows:

$$Uplift(x_i) = P(Y = 1|x_i; t_i = 1) - P(Y = 1|x_i; t_i = 0) \quad (1)$$

As a formal definition, let X be a vector of predictor or independent variables, $X = \{x_1, \dots, x_m\}$ and $Y\epsilon\{0,1\}$ be the binary dependent class variable whose behavior is to be modeled. $Y = 1$ is assumed to be the positive outcome (success), and $Y = 0$, negative. In addition, $T\epsilon\{0,1\}$ represents whether or not a given object is in the treatment group, $T = 1$, or in the control group, $T = 0$. Finally, $P$ denotes a probability as predicted by a model. Figure 1 provides a conceptual overview of uplift modeling.

*Two-Model Approach*

This approach builds on the traditional classification models and consists of two separate predictive models $M_T$ and $M_C$, using the treatment group data and the control group data, respectively. Subtracting the estimate from $M_C$ from the estimate from $M_T$ gives the final uplift.

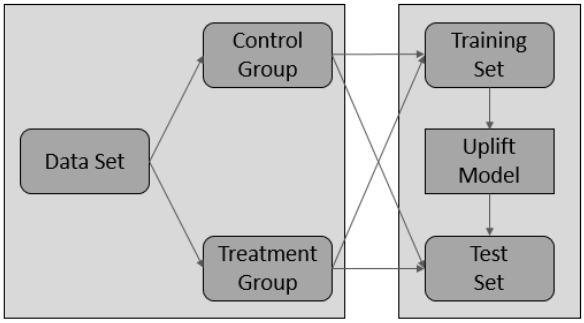
$$\begin{aligned} M_{Uplift} &= M_T - M_C \\ &= P(Y = 1|x_i; t_i = 1) - P(Y = 1|x_i; t_i = 0) \end{aligned} \quad (2)$$

The main advantage of this approach is its simplicity. Its implementation is straightforward, allowing users to use any state-of-the-art machine learning algorithms, such as logistic regression or support vector machines for building the treatment and the control models.

In contrast, the major drawback of this approach is its limited applications in the simplest cases or, in other words, its practical failure in real-world situations (Radcliffe 2011). One reason for this is both the treatment and the control models are built independently and they focus on predicting the outcome separately without taking into account one another and they do not explicitly estimate the uplift. This approach could also result in having different sets of predictor variables for each model because of the independent construction of the models. For example, each model can prioritize different variables that are having high predictive power to estimate the outcome while disregarding the variables that best estimate the uplift across two models.

**Figure 1**

The general representation of uplift modelling



The two-model approach is an indirect means of estimating uplift and as a result, having associated problems and limitations. However, these can be overcome by building a single model that directly predicts the uplift using the entire dataset without separating it into the treatment and the control groups.

*Dummy Treatment Approach*

This approach combines the treatment and the control groups into a development sample and creates a dummy treatment variable in that sample. Based on logistic regression, it estimates the uplift via variables and their interactions with the dummy treatment variable, $T$ (Lo 2002). It will assign $T = 1$ for the observations from the treatment group and $T = 0$ for the control group. This method consists of predictor variables $X$ capturing the baseline estimates for the control group, treatment variable $T$, and interaction variables $X \cdot T$ capturing the uplift estimate:

$$P_i = E(Y_i|X_i) = \frac{\exp(\propto + \beta' X_i + \delta T_i + \gamma' X_i T_i)}{1 + \exp(\propto + \beta' X_i + \delta T_i + \gamma' X_i T_i)} \tag{3}$$

$$Uplift_{Lo}(x) = P_i|treatment - P_i|control$$

$$= \frac{\exp(\propto + \gamma + \beta' X_i + \delta' X_i)}{1 + \exp(\propto + \gamma + \beta' X_i + \delta' X_i)} - \frac{\exp(\propto + \beta' X_i)}{1 + \exp(\propto + \beta' X_i)} \tag{4}$$

where $\propto, \beta, \gamma$, and $\delta$ are parameters to be estimated.

While the interaction variables make this model account for the heterogeneity in the treatment response originated by patients' characteristics, they may also introduce collinearity issues into the model. These variables, for instance, are used both as baseline and interaction variables. This can result in instability and overfitting (Kane 2014).

*Transformation Approach*

This approach was first proposed by Lai (2006) for customer classification in direct marketing by using outcome variable transformation. In direct marketing, the customer population is generally stratified into four groups on whether a customer responds when treated or not (Figure 2).

**Figure 2**

The categorization of observations based on whether an individual was treated and whether the individual responded.

**From four known outcomes**

| **Treated** | **Responded: Yes** | **Responded: No** |
|---|---|---|
| **No** | Control Responders (CR) | Control Non-Responders (CN) |
| **Yes** | Treated Responders (TR) | Targeted Non-Responders (TN) |

Lai's method labels control non-responders (CN) and treated responders (TR) as positive targets while labeling control responders (CR) and targeted non-responders (TN) as negative targets. Any supervised classification technique can be used to estimate the outcome probabilities of these four quadrants. Labeling these quadrants transforms the target variable into a binary target variable, and as a result, it converts the uplift modeling into a binary classification problem. Hence, the uplift of treatment can be calculated as follows by subtracting the probability of negative targets from the positive ones:

$$Uplift_{Lai}(x) = [P(TR|x) + P(CN|x)] - [P(TN|x) + P(CR|x)] \tag{5}$$

Kane et al. (2014) have proved that Lai method is mathematically correct only when both the treatment and the control groups have the same size of observations and are randomly selected. Imbalance group sizes might introduce bias to the estimated probabilities. In practice, randomized experiments and RCTs most often have imbalanced treatment and control observations and this limits the application areas of this approach to a great degree.

By weighing the probability scores of each quadrant, Kane et al. (2014) correct or generalizes the Lai's method and recognizes the impact of the treatment-to-control ratio on the estimated uplift. The proposed equation for the uplift is as follows:

$$Uplift_{Generalized\ Lai}(x) = \frac{P(TR|x)}{P(T)} + \frac{P(CN|x)}{P(C)} - \frac{P(TN|x)}{P(T)} + \frac{P(CR|x)}{P(C)} \tag{6}$$

where $P(T)$ represents the proportion of individuals in the treatment group and $P(C)$ the proportion of individuals in the control group, and $P(C) = 1 - P(T)$.

A similar transformation approach was also proposed by Shaar et al. (2016). Their approach is as follows: they develop re-weighted Lai's approach (7), at the same time; they calculate the reflective uplift (8). The proposed reflective uplift predicts the probability of an individual being treated given the individual has responded by applying a two-model approach, $M_R$ for the responders and $M_N$ for the non-responders. Finally, they combine the reflective uplift and re-weighted Lai's approach in order to get the pessimistic uplift (9). They state that their model minimizes the effect of the noise in the data by using the reflective uplift as a stabilizer.

$$\begin{aligned} Uplift_{Re-weighted\ Lai}(x) = {} & [P(TR|x) + P(CN|x)] * P\left(\frac{TR + CN}{population}\right) \\ & -[P(TN|x) + P(CR|x)] * P\left(\frac{TN + CR}{population}\right) \end{aligned} \tag{7}$$

$$Uplift_{Reflective}(x) = P_{Reflective}(Positive|x) - P_{Reflective}(Negative|x) \tag{8}$$

$$P_{Reflective}(Positive|x) = P_{M_R}(T|R) * P(TR) + P_{M_N}(C|N) * P(CN)$$

$$P_{Reflective}(Negative|x) = P_{M_N}(T|N) * P(TN) + P_{M_R}(C|R) * P(CR)$$

$$Uplift_{Pessimistic} = \frac{1}{2} * (Uplift_{Re-weighted\ Lai} + Uplift_{Reflective}) \tag{9}$$

The advantage of the Lai and its generalized version is that it converts an arbitrary probabilistic classification model into a single model that directly predicts the uplift. Thus, it allows users to use traditional supervised classification techniques. Whereas, the pessimistic uplift provide a more stable model.

### 2.2. Literature Review

Many studies have been conducted to investigate the effect of HTM on patients with HF compared to standard care. These RCTs demonstrated that HTM intervention reduced the risk of all-cause mortality, heart failure-related hospitalizations, and improvements in health-related quality of life (Cleland 2005, Koehler 2018, Inglis 2015)(Koehler 2018, Inglis 2015, Cleland 2005). In those trials, authors mainly propose regression analysis based approaches for modeling the difference between the treatment and the control groups. Although they are very insightful, the purpose of those methods is different from the problem discussed here. Those studies are not addressing the subject heterogeneity; they only focus on the overall effect of the treatment across the full population of interest. Whereas, the main purpose of this thesis is to define the subgroup of HF patient in which the treatment be most beneficial.

A large number of existing literature on uplift modeling concentrated mainly on the field of direct marketing (Coussement 2017, Devriendt 2018, Lo 2015, Marinakos 2017, Rzepakowski 2012a). Because of the drawbacks of the two model approaches as reported by Radcliffe et al. (2011), there are several studies proposing direct modeling of the uplift using different approaches. Logistic regression with transformation approach (Kane 2014, Rudaś 2018, Lai 2006), Support Vector Machines (Zaniewicz 2013), and k-Nearest Neighbors (Berezin 2015). Additionally, there are other uplift modeling studies adapting decision trees to split treatment group data (Radcliffe 2011, Rzepakowski 2012b), while there are also other studies combining decision trees into ensemble methods (Guelman 2014, Sołtys 2015).

Despite its practical importance in the medical context, there has been limited attention to the uplift modeling for patient stratification for medical treatments. Jaroszewicz et al. converted the uplift modeling problem into a binary classification problem by using a transformation approach (2012). They tested this approach on three publicly available datasets from R statistical system packages. The first dataset covers patients who received two types of a bone marrow transplant, the second focuses on the treatment of breast cancer with tamoxifen, and the last one consists of survival of patients with hepatitis. The performance of the method was not satisfactory and was outperformed by two model approach. Later, Rzepakowski together with Jaroszewicz proposed decision tree construction for uplift modeling and tested this method with the above-mentioned datasets (2012b). The authors modify the splitting criteria and a tree pruning for the uplift modeling case and demonstrated significant improvement in the performance of the model. This method further combined into ensemble methods by Sołtys et al., and their experiments on the same datasets showed performance improvements (2015). The same authors also applied a variety of boosting algorithms with uplift modeling and showed that AdaBoost applicability with the same medical datasets (2015b).

Finally, the above-mentioned studies demonstrated the performances of their uplift models on publicly available datasets which are available in various R packages. There are studies using

RCT datasets directly from universities and clinics, as well. Chang et al. (2019) presented successful patients subgrouping for whom ambulatory surgical cleft lip repair is more likely to be beneficial. The researchers applied uplift modeling with two model approach based on logistic regression in their study. Moreover, logistic regression based uplift modeling was demonstrated using three uplift modeling approaches, mentioned in Section 2.1, by Biswas et .al (2018) to stratify patients with regard to electronic alert for a cute kidney injury.

## 3. Experimental Setup

### 3.1. Data

We investigated and evaluated various methods of uplift modeling on two benchmark datasets. One of them is a real-world dataset from a randomized clinical trial testing an intervention in a heart failure patient cohort and the other one is a set of generated data comprising an intervention and control in synthesized heart failure patient cohort. Table 1 summarizes the main characteristics of these datasets.

The clinical trial data is from a previously published randomized trial of the remote monitoring treatments on patients with heart failure, TEN-HMS study (The Trans-European Network-Home-Care Management System) (Cleland 2005). Briefly, 426 patients with heart failure were assigned randomly to receive home telemonitoring (HTM), nurse telephone support (NTS), and usual care (UC) in a 2:2:1 ratio. The main comparison of interest in our project is between HTM and UC groups. The UC group was used as a control group to predict the incremental benefits of the HTM treatment on the subgroup of the patients. The trial took place in Germany, the Netherlands, and the United Kingdom over a median follow-up of 484 days. There are 118 features in the dataset. All of the outcome features were created using related features from the dataset, except for the vital status outcome. Table 2 provides the main characteristics of these features. The "signal-to-noise ratio" is the difference between the positive respond rates of treatment and control groups over that of the control group (Kane 2014). The positive response rate refers to the ratio of the positive (favorable) outcome observations over negative (unfavorable) ones. We defined the outcome features as follows:

**Table 1**
The overview of datasets used in the experiments.

| **Variable** | TEN-HMS dataset | Synthetic dataset |
|---|---|---|
| No. of observations | 242 | 24200 |
| No. of treatment observations | 162 | 15272 |
| No. of control observations | 80 | 8928 |
| No. of variables | 88 | 26 |
| Treatment-to-control size ratio | 2.0:1 | 1.7:1 |

Binary outcome features

- *Vital Status*:
  Participants who were still alive during the trial were assigned with favorable outcome value, 1, while the ones who were lost due to all-cause death were assigned with unfavorable outcome value, 0.
- *Death or all-cause hospitalization*:
  Participants who were still alive and did not readmit to hospital during the trial were assigned with favorable outcome value, 1, while the ones who were lost due to all-cause death or readmitted to the hospital at least once because of any cause were assigned with unfavorable outcome value, 0. Any first hospital readmission within the 30 days of the patient's enrolment to the trial was neglected because HTM or UC treatments have no effect on those readmissions.
- *Death or heart failure-related hospitalization*:
  Participants who were still alive and did not readmit to the hospital because of HF during the trial were assigned with favorable outcome value, 1, while the ones who were lost due to all-cause death or readmitted to the hospital at least once because of HF was assigned with unfavorable outcome value, 0. Any first hospital readmission within the 30 days of the patient's enrolment to the trial was neglected because HTM or UC treatments have no effect on those readmissions.
- *240-day all-cause hospitalizations*:
  Participants who did not readmit to the hospital within the 240 days after their enrolment to the trial were assigned with favorable outcome value, 1, while the ones who readmitted to the hospital due to any cause in this period were assigned with unfavorable outcome value, 0. Any first hospital readmission within the 30 days of the patient's enrolment to the trial was neglected because HTM or UC treatments have no effect on those readmissions.

**Table 2**
The overview of the outcome features used in the experiments.

| | TEN-HMS dataset | | | Synthetic dataset | | |
|---|---|---|---|---|---|---|
| Outcome Features | Treatment positive rate | Control positive rate | Signal-to-noise ratio | Treatment positive rate | Control positive rate | Signal-to-noise ratio |
| Vital Status | 74.70% | 62.50% | 19.52% | 70.90% | 56.64% | 25.18% |
| Death or all-cause hospitalization | 37.60% | 32.50% | 15.69% | 42.10% | 34.90% | 20.63% |
| Death or heart failure-related hospitalization | 61.10% | 56.20% | 8.72% | 57.50% | 51.07% | 12.59% |
| 240-day all-cause hospitalizations | 51.20% | 43.70% | 17.16% | 55.90% | 49.26% | 13.48% |

The outcome feature creations were all done before performing any experiments in order to avoid biasing the results. During the experiments, we omitted all of the features that were used for creating outcome features in order to prevent any data leakage. One of the reasons for the data leakage is when the dataset used for the training of a machine learning model happens to have one or more features that are inherently proxy for the outcome.

The simulated data was generated using the Synthpop package in the R statistical system. We generated 100 datasets and combined them together for our experiments. The generated datasets mimic the original TEN-HMS dataset and preserve the relationship between the features. Especially, the distribution of selected outcome features was estimated conditional on all of the predictor features. Further details about the package features are described by Nowok et al. (2016).

### 3.2. Method

We describe the mathematical formulation of the uplift models that we used during this project in Section 2. If there is a randomized control group in a dataset, then modeling the uplift in a study is straightforward. One can calculate the uplift predictions for different uplift techniques according to equations provided in Section 2.2. The full overview of the approaches that we implemented during this study is shown in Table 3**.** On the other hand, assessing the performance of an uplift model is more complex than assessing more conventional machine learning models. In practice, the standard evaluation for a traditional model is to use cross-validation. Cross-validation can be explained as: partitioning a part, generally 80 percent, of the dataset into training and validation sets, training the model on the training set, predicting the targets on the validation set, and finally validating the model's performance by comparing it to the ground truth (the target values of the testing set). In uplift modeling, although we can still use the cross-validation, we cannot validate the predicted results by comparing them to the ground truth. The predictions of the uplift models are probability differences between two groups. There is no ground truth for the predictions of uplift models due to the Fundamental Problem of Causal Inference which was mentioned in Section 2.1. For a given individual, we cannot observe the effect of being treated or not treated at the same time. Subsequently, the uplift, or the target variable, on a given individual is not observable; we cannot directly calculate the error of an uplift model by comparing the predicted outcomes at a level of a single identity.

Due to the aforementioned issues related to the fundamental problem of causal inference, we cannot use only an independent test set to evaluate uplift models; different evaluation measures and visual evaluation approaches are necessary. The first step for the evaluation is randomly splitting the dataset into training and test datasets. A related point to consider is keeping the same distribution of treatment and control group observations, and also the distribution of the outcome

**Table 3**
The overview of the outcome features used in the experiments and used hyperparameters.

| Uplift modeling approach | Code | Classifier | Performance metric |
|---|---|---|---|
| Two-model approach | *tma* | Logistic regression | Receiver Operating Characteristics |
| Dummy treatment approach | *dta* | Logistic regression | Receiver Operating Characteristics |
| Generalized Lai's approach | *glai* | Stochastic gradient boosting | Receiver Operating Characteristics |
| Pessimistic approach | *pess* | Logistic regression | Receiver Operating Characteristics |

Hyperparameter settings for the stochastic gradient boosting

| Max tree depth | No. of boosting iterations | Min. Terminal Node Size | Shrinkage |
|---|---|---|---|
| 1 | 50 | 10 | 0.1 |

feature within these subsets during random dataset partitioning in order to prevent possible bias sources. Consequently, the following evaluation approaches can be applied:

*Uplift by Segment*
For each observation in the test set, we compute the predicted uplift scores. As described in Section 2, the predicted uplift scores can be calculated according to equations (2), (4), (6) and (9); for a two-model (*tma*), dummy treatment (*dta*), transformation (generalized Lai, *glai* and pessimistic, *pess*) approaches, respectively. Afterward, we rank all of the observations from both the treatment and control groups in descending order and group them in segments of equal observations number. Here the number of segments is arbitrary and should be chosen according to the size of the test set. For each segment, we calculate the average scores of the treatment and the control groups and take the differences between them to calculate the actual uplift. The comparison between treatment and control in each segment is based on an assumption that individuals from the treatment group who were assigned similar uplift scores are actually having a counterpart in the control group. Presenting the predicted score differences per segment side by side in a bar plot, we can have an idea of the uplift per segment and visually asses how the model performs.

*Qini R-squared metric*
This metric is the R-squared ($R^2$) of a regression line fitted on semi-segment values of the uplift by segment chart. An ideal uplift model yields an uplift by segment chart that displays uplift scored segments in descending order from left to right with smoothly declining uplift scores per segment, which gives the highest R-squared value of 1 for the fitted regression line. However, in practice, most uplift models yield a chart with unevenly declining segments, which, as a result, decreases the R-squared value. Therefore, Qini R-squared metric can be used as a secondary

metric to compare the smoothness of the declining segment sequences of the uplift be segment charts.

*Qini Curves*

Qini curve is an alternative visual performance assessment for uplift models. Qini curve, or cumulative uplift, was introduced by Radcliffe (2007). In order to obtain the Qini curve, we subtract the cumulative uplift scores of the control group from the treatment group for each segment in the test set. Next, we plot these cumulative uplift score differences as a function of the fraction of the individuals treated; from no one treated to full population treated. Before the plot, the different fraction of individuals is sorted by the predicted uplift in the descending order. The Qini curve can be formulated as:

$$Q(t) = R_t^T - \frac{R_t^C N_t^T}{N_t^C} \tag{10}$$

where the $t$ subscript indicates first $t$ observations. $R_t^T$ and $R_t^C$ are the sum of the predicted uplift scores, and $N_t^T$ and $N_t^C$ are the total number of observations in the segment for the treatment and the control groups, respectively. It is important to note that Qini curves evaluate uplift performance by comparing groups of individuals rather than single individuals. Qini curves allow us to identify the possible highest uplift by a model, and the uplift for any segment.

*Qini Measures*

Although the above-mentioned uplift evaluation methods are useful, they do not provide any means to compare different models accurately. Qini measure, on the other hand, is the most detailed and direct measure. It was introduced by Radcliffe (2007) and similar to the area under the uplift curve (AUUC) (Rzepakowski 2010). It is adapted from the Gini measures from economics (Lerman 1984). The Qini measure is the area between the Qini curve of the uplift model and the diagonal line of the random targeting.

*Top 20% Qini Measure*

It is Qini measure value for the top 20 percent of uplift scores. Qini measure evaluates uplift model performance over the full population. However, the aim of uplift modeling is to rank the population and identify a subgroup of individuals who are expected to have an increased effect from the treatment. For our case, we are interested in the observations with predicted uplift scores that are in the 20% highest ranked. Here, the top percent value is arbitrary and it can be modified for any given application.

All of the experiments were conducted in the open-source statistical software `R` together with the `caret` package (R Development Core Team 2016, Kuhn 2017).

## 4. Results

In this section, we present the uplift model applications to the problem of identifying subgroups of patients that receive benefit from an HTM intervention based on the datasets described in Section 3.1. In all of the experiments, we compare the performances of two-model (*tma*), dummy treatment (*dta*), generalized Lai (*glai*), and pessimistic (*pess*) approaches.

### 4.1. TEN-HMS Dataset

As can be seen from Table 1 in Section 3, TEN-HMS dataset includes many variables. If a high number of variables are used as predictor variables in an uplift model, this may cause overfitting due to correlations between the predictor variables and an increase in complexity of the model (García 2015). Compared to the high number of variables, the number of observations in TEN-HMS dataset is quite low. To reduce the number of variables that can run as a predictor candidate in the model, we have performed the variable pre-selection using subject knowledge by consulting experts in the field. A list of 17 candidate predictor variable was selected from a total of 88 variables for the experiments.

It is also worth mentioning that the evaluation of the models was performed using TEN-HMS dataset without any data splitting for hold-out or cross-validation. We have conducted a self-test in which the test set to evaluate the uplift models is exactly the same as the training set to train the uplift models. Considering an 80 percent split of the dataset for the model validation, the test sets would only contain 48 observations after the data splitting. This number of observations is not sufficient to have a proper uplift model evaluation. This is further discussed in Section 5.

The performances of four approaches in terms of the Qini measure, Top 20% Qini measure, and the Qini R-squared metric are reported in Table 4 for TEN-HMS dataset. Comparison of the model's performances with respect to their Qini measures is shown in Figure 3. From the results, it is clear that generalized Lai' approach (*glai*) outperforms other approaches across all of the outcome features. It generates higher values for all of the three Qini evaluation measures. Additionally, *tma* and *dta* approaches produce identical results across all of the outcome features. Although the *pess* approach yields higher scores in Qini measures for all of the outcomes, its performance is below the performance of the baseline model *tma*, with respect to the Top 20% Qini measure.

The effect of the number of predictor variables on the model performance in terms of the Qini measure is shown in Figure 4. For this purpose, we used the outcome feature with the highest Qini value, "*Death or heart failure-related hospitalization*". As shown in Figure 4, the best performance is achieved when the highest number of predictor variables is used in the model; that is 17 features for the *glai* method. The performance of the *pess* approach shows a descending trend after the addition of 14 variables, whereas this descending trend starts after the addition of 11 variables for both *tma* and *dta* approaches. For both of *glai* and *pess* methods, the performances increase slightly with the use of 17 features.

We perform visual evaluation techniques to have further details on the performances of the models and for their comparisons. Figure 4 lists the uplift by segment and the Qini curve charts of all techniques with the above-selected outcome, "*Death or heart failure-related hospitalization*".

The uplift by segment chart can be interpreted as follows: the leftmost bar corresponds to the uplift in the first segment, in this case, quintile; the subsequent bar corresponds to the first 40 percent, and so on. As described in Section 3.2, the number of segments should be chosen according to the size of the dataset. For larger datasets, deciles are generally chosen as the segments. The R-squared of a fitted straight line on the uplift by segment chart gives an estimate for the model repeatability, which is Qini R-squared value. The ideal chart displays uplift scored segments in descending order from left to right, resulting in the highest R-squared value of one. As there is a direct association between the Qini curve and the uplift by segment chart, the higher bars in the chart generates higher Qini curve above the diagonal line of the random targeting.

We can interpret the Qini curve as follows: the x-axis represents the percentage of the population on which treatment is performed, and the y-axis shows the cumulative uplift difference. The 100% on x-axis gives the estimated cumulative uplift if the entire population would be treated. The slope of the diagonal line of random targeting gives an idea about the impact of the treatment on the entire population. The positive slope of this line implies an overall beneficial effect of the treatment when the entire population would be targeted.

As can be seen from the Figure 5, the outcome feature, "*Death or heart failure-related hospitalization*", achieve the highest uplift of approximately 29% with *glai* method when the treatment is applied on 60% of the population, while an uplift of approximately 5% is achieved when the entire population is treated. Additionally, with the *glai* method, more symmetrical uplift by segment chart is generated with a high Qini R-squared value of 0.866 (Table 1). Both *tma* and *dta* approaches generate only around 17% uplift when 60 percent of the population is treated.

**Figure 3**

The effect of predictive variable number on the uplift models with the outcome feature of "*Death or heart failure-related hospitalization*" in TEN-HMS dataset

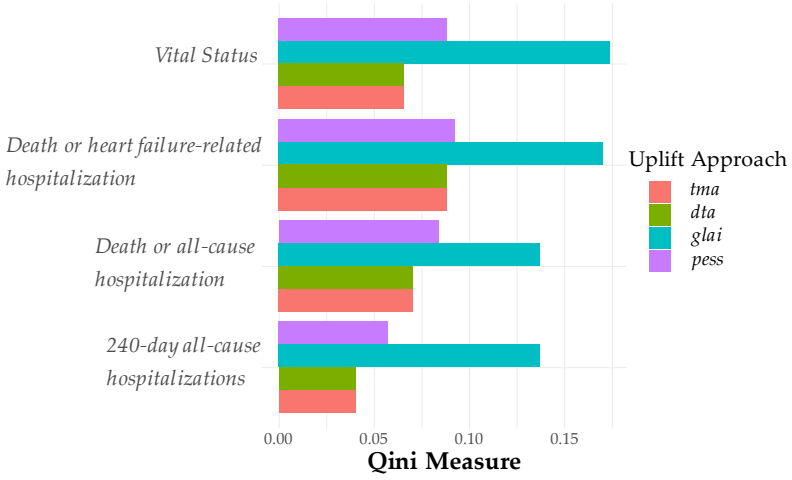


**Table 4**
Model Performance for each approach across the outcome features on TEN-HMS dataset

| | *tma* | | | *dta* | | |
|---|---|---|---|---|---|---|
| Outcome Feature | Qini | Top 20% Qini | Qini R-squared | Qini | Top 20% Qini | Qini R-squared |
| Vital Status | 0.066 | 0.0093 | 0.836 | 0.066 | 0.0093 | 0.836 |
| Death or all-cause hospitalization | 0.0707 | 0.01 | 0.891 | 0.0707 | 0.01 | 0.891 |
| Death or heart failure-related hospitalization | 0.088 | 0.0065 | 0.759 | 0.088 | 0.0065 | 0.759 |
| 240-day all-cause hospitalizations | 0.0407 | 0.0066 | 0.773 | 0.0407 | 0.0066 | 0.773 |

| | *pess* | | | ***glai*** | | |
|---|---|---|---|---|---|---|
| Outcome Feature | Qini | Top 20% Qini | Qini R-squared | Qini | Top 20% Qini | Qini R-squared |
| Vital Status | 0.088 | 0.009 | 0.871 | 0.1737 | 0.0145 | 0.988 |
| Death or all-cause hospitalization | 0.0838 | 0.0069 | 0.819 | 0.137 | 0.0132 | 0.93 |
| **Death or heart failure-related hospitalization** | 0.0923 | 0.0058 | 0.674 | **0.1701** | **0.0154** | **0.866** |
| 240-day all-cause hospitalizations | 0.0573 | 0.0034 | 0.65 | 0.1369 | 0.0124 | 0.916 |

Using the *glai* approach, we map the characteristics distribution of the patients who will benefit most from the treatment (Figure 6). It appears that the patient's characteristics as NYHA (New York Heart Association), BMI (Body Mass Index), a number of prior hospitalizations, and ejection fraction levels can be used together in the recruiting process of patients with heart failure for an HTM treatment. From the distribution, it can be seen that patients with NYHA level of two (mild symptoms with mild shortness of breath and/or angina, and slight limitation during ordinary activity), having BMI around 22 (healthful weight), less than 5 prior hospitalization, together with the ejection fraction level of around 20 (pumping ability of the heart is severely below normal) will most likely get the highest benefit if they will be recruited into an HTM treatment.

### 4.2. Synthetic Dataset

As described in Section 3.1, the synthetic dataset is generated to mimic the TEN-HMS dataset, consisting of the same 17 predictive variables and outcomes features from the TEN-HMS dataset. We evaluated all four uplift approaches by applying fivefold stratified cross-validation with a fixed seed in order to rerun experiments and validate results. To ensure consistency in model comparisons, the same test set was used for obtaining figures and tables.

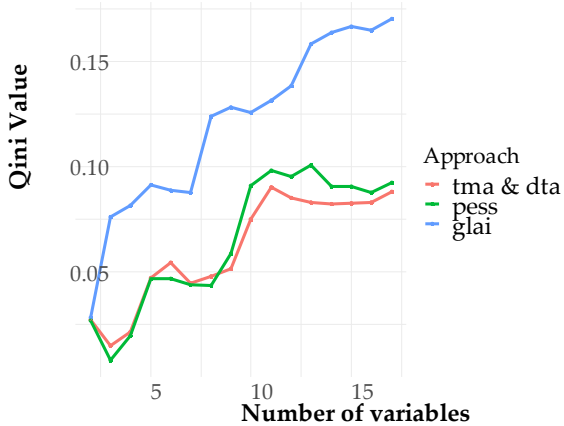
Table 5 shows the performances of four approaches in terms of the Qini measure, Top 20% Qini measure, and the Qini R-squared metric. The results show that *glai* outperforms other approaches across all of the outcome features with respect to all three uplift evaluation metrics. The highest Qini and Top 20% Qini are achieved with the outcome feature "*Death or heart failure-related hospitalization*". *tma* and *dta* approaches produce identical results across all of the outcome features. The *pess* approach performs worst than our baseline model *tma*. When compared to TEN-HMS dataset, the model performances decrease for the synthetic dataset. This decrease is further discussed in Section 5.

For the synthetic dataset, we compare the robustness of the model performances using the same three metrics as used previously. We assess the robustness by applying the models on five different random sample sets of the synthetic dataset and the entire dataset. Each sample set consist of 50 percent of the synthetic dataset. For this experiment, we choose the outcome feature of "*Death or heart failure-related hospitalization*", with which the highest metric values are achieved (Table 5). Table 6 depicts the predicted metric values from this experiment. The results confirm that for the *glai* approaches, the Qini measure shows little variance. The little standard deviation in all three metrics highlights the stability of the *glai* model. For each outcome feature and uplift approach, we used the same sample sets of the synthetic dataset, which are randomly taken from the dataset while keeping the same distribution of treatment and control group observations, and outcome feature distribution within these subsets.

Figure 7 lists the uplift by segment and the Qini curve charts of all techniques with the above-selected outcome feature. The highest uplift of approximately 12% is achieved with *glai* method when the treatment is applied on 60 percent of the population, while an uplift of approximately

**Figure 4**

The effect of predictive variable number on the uplift models with the outcome feature of "*Death or heart failure-related hospitalization*" in TEN-HMS dataset



9% is achieved when the 20% of the population is treated. The *glai* method generates an uplift by segment chart which is declining smoothly across the deciles with the Qini R-squared metric of 0.835. On the other hand, both *tma* and *dta* approaches, which performed better than the *pess* approach, generate only around 6% uplift when 70 percent of the population is treated. The uplift by segment chart generated by *tma* and *dta* has the top deciles lower than the 6$^{th}$ and 8$^{th}$ deciles which are interrupting the declining trend of the deciles. Therefore, we can conclude that these two approaches yielding the chart are not performing well and have limited practical value for ranking the patients for the synthetic dataset.

In our model evaluations of the two datasets, generalized Lai's approach appears to be the best model with generating metrics that are outperforming the other models across all of the outcome features. Especially, the Top 20% Qini metrics from the *glai* approach are all above the diagonal line of the random targeting and the other models corresponding values. From a practical perspective, it is more motivating to reach highest uplift scores on the smaller fraction of the population. The aim of the uplift modeling is to rank and subgroup the population for which the treatment is expected to have an increased effect.

Using the *glai* approach, we also map the characteristics distribution of the patients who will benefit most from the treatment (Figure 8). From the distribution, it can be seen that patients with NYHA level of one (no symptoms and no limitation in ordinary physical activity), having BMI around 22 (healthful weight), together with less than 5 prior hospitalizations will most likely get the highest benefit if they will be recruited into an HTM treatment.

**Figure 5**

Uplift by segments charts and Qini curves from TEN-HMS dataset: (a) two-model approach (*tma*) and dummy treatment approach (*dta*); (c) pessimistic approach (*pess*); (c) generalized Lai approach (*glai*) with the outcome feature of "*Death or heart failure-related hospitalization*"

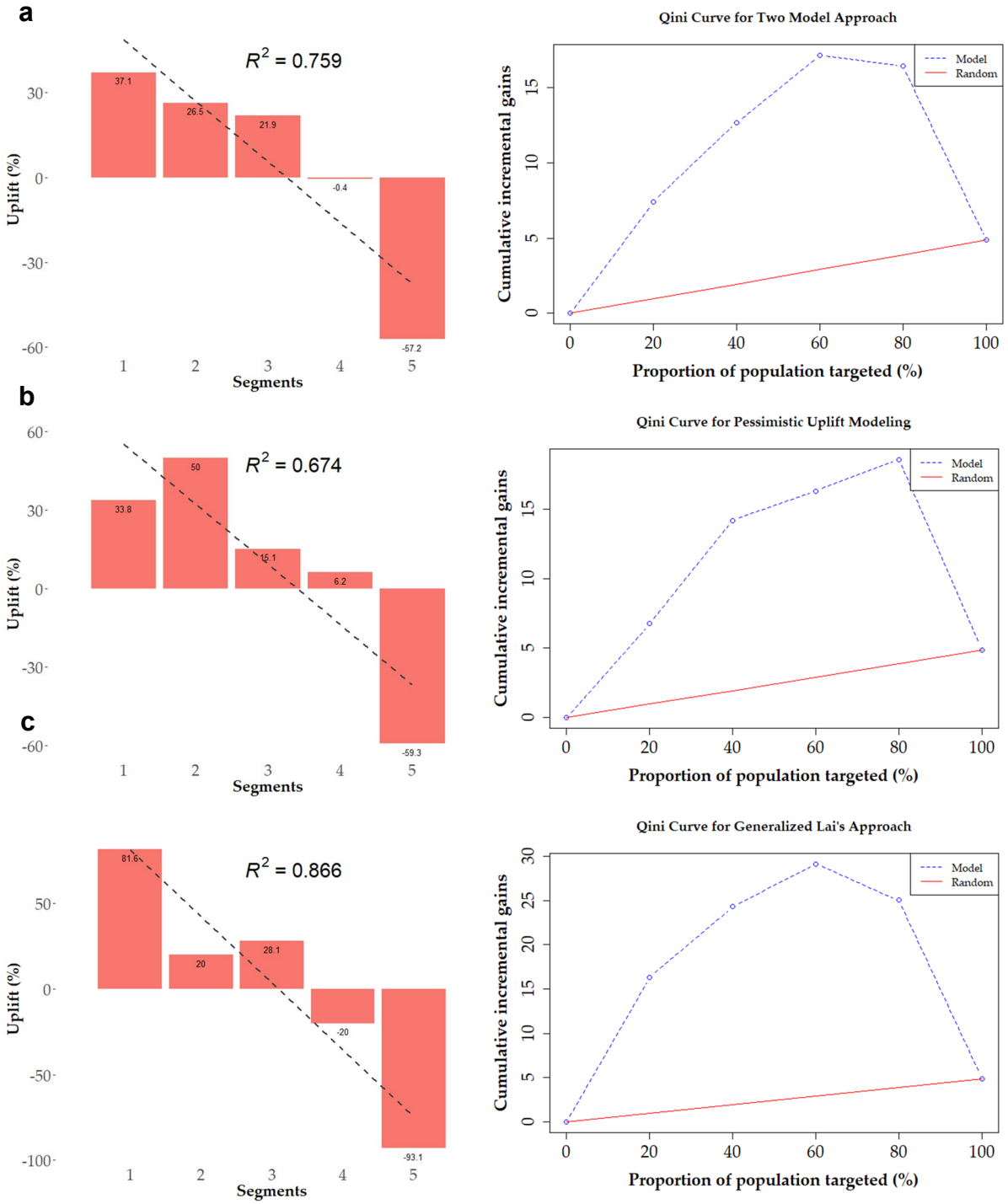


**Table 5**
Model Performance for each approach across the outcome features on the synthetic dataset

| | *tma* | | | *dta* | | |
|---|---|---|---|---|---|---|
| Outcome Feature | Qini | Top 20% Qini | Qini R-squared | Qini | Top 20% Qini | Qini R-squared |
| Vital Status | 0.0066 | 0.0006 | 0.283 | 0.0066 | 0.0006 | 0.283 |
| Death or all-cause hospitalization | 0.0257 | 0.0013 | 0.561 | 0.0257 | 0.0013 | 0.561 |
| Death or heart failure-related hospitalization | 0.0138 | 0.0012 | 0.371 | 0.0138 | 0.0012 | 0.371 |
| 240-day all-cause hospitalizations | 0.0264 | 0.0028 | 0.703 | 0.0264 | 0.0028 | 0.703 |

| | *pess* | | | ***glai*** | | |
|---|---|---|---|---|---|---|
| Outcome Feature | Qini | Top 20% Qini | Qini R-squared | Qini | Top 20% Qini | Qini R-squared |
| Vital Status | 0.003 | 0.0007 | -0.116 | 0.0553 | 0.0074 | 0.68 |
| Death or all-cause hospitalization | 0.0016 | 0.0002 | -0.034 | 0.0507 | 0.0033 | 0.811 |
| **Death or heart failure-related hospitalization** | 0.0113 | 0.0020 | 0.814 | **0.0652** | **0.0084** | **0.835** |
| 240-day all-cause hospitalizations | 0.0097 | 0.0008 | 0.079 | 0.0619 | 0.0053 | 0.857 |

**Figure 6**
Characteristics distribution of the patients from TEN-HMS dataset; the top 20 percent segment in which patients will benefit most from the treatment, and the bottom 20 percent segments in which patients will benefit less from the treatment

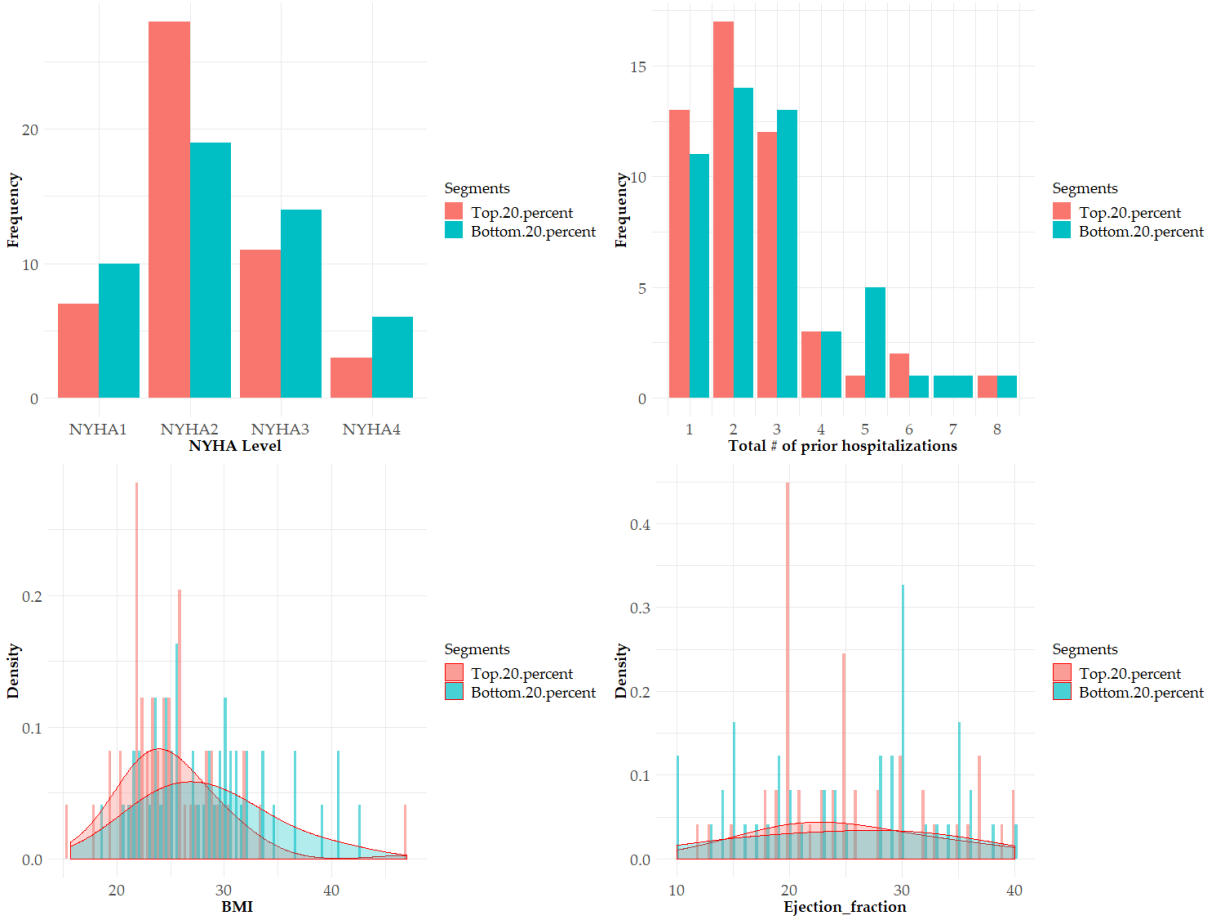


**Figure 7**
Uplift by segments charts and Qini curves from synthetic dataset: (a) two-model approach (*tma*) and dummy treatment approach (*dta*); (b) pessimistic approach (*pess*); (c) generalized Lai approach (*glai*) with the outcome feature of "*Death or heart failure-related hospitalization*".

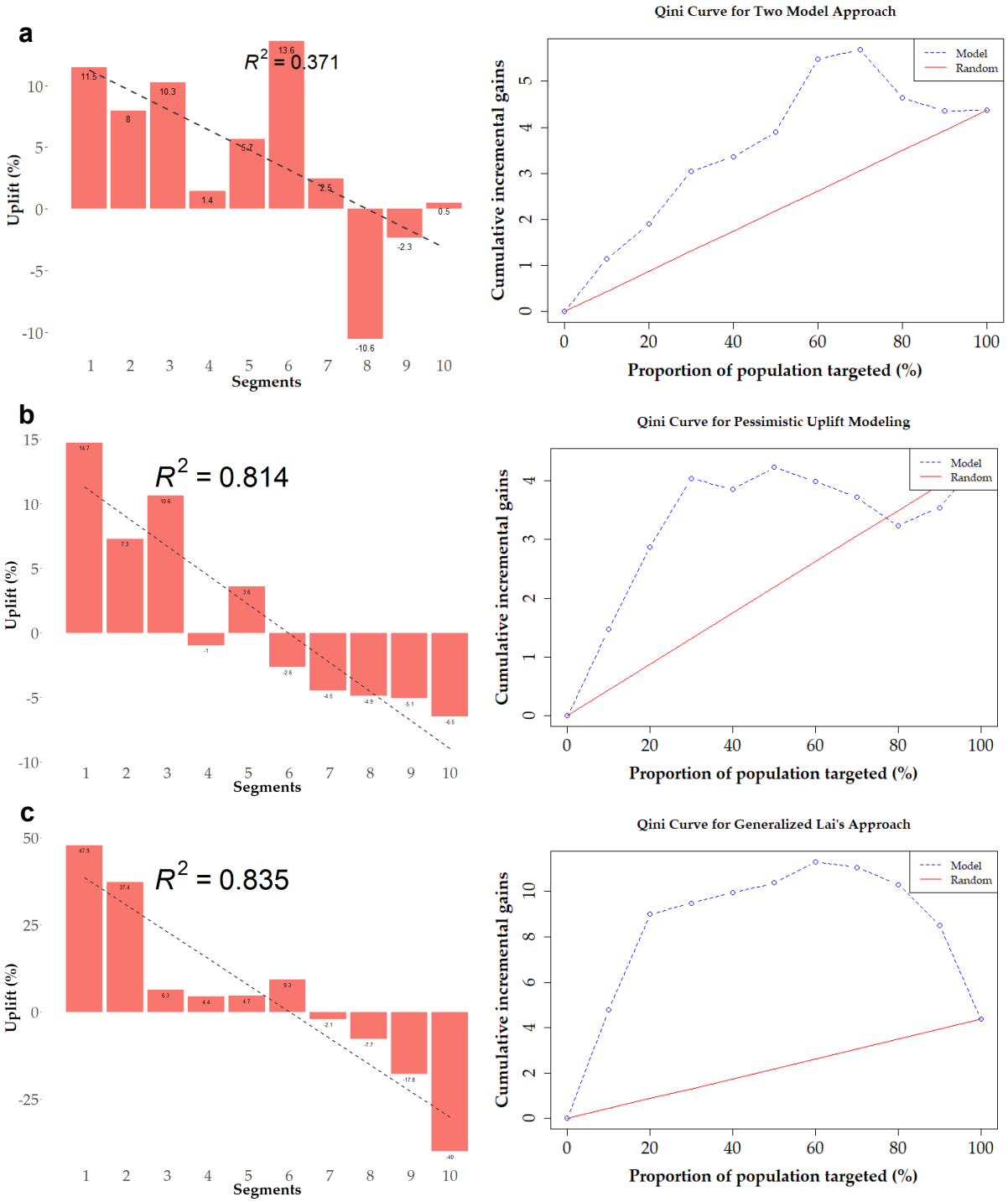
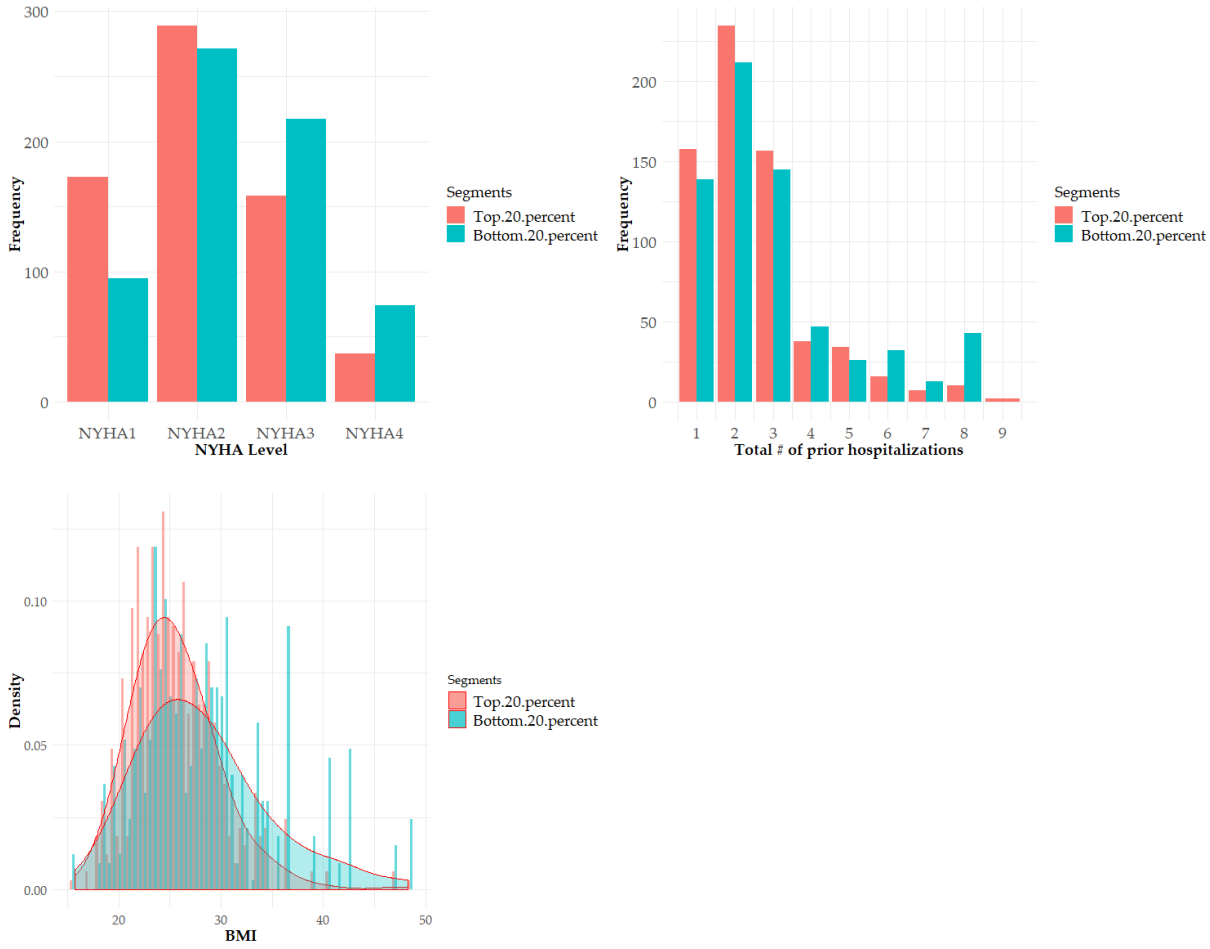


**Table 6**
Stability of different approaches achieved with the outcome feature "*Death or heart failure-related hospitalization*" on 50% samples and the full synthetic dataset

| | *tma* | | | *dta* | | |
|---|---|---|---|---|---|---|
| Data Sample | Qini | Top 20% Qini | Qini R-squared | Qini | Top 20% Qini | Qini R-squared |
| 50% Sample 1 | 0.0202 | 0.0014 | 0.765 | 0.0202 | 0.0014 | 0.765 |
| 50% Sample 2 | 0.0167 | 0.0029 | 0.282 | 0.0167 | 0.0029 | 0.282 |
| 50% Sample 3 | 0.0185 | 0.0014 | 0.475 | 0.0185 | 0.0014 | 0.475 |
| 50% Sample 4 | 0.0052 | 0.0003 | -0.125 | 0.0052 | 0.0003 | -0.125 |
| 50% Sample 5 | 0.0225 | 0.0033 | 0.474 | 0.0225 | 0.0033 | 0.474 |
| **Full Sample (100%)** | **0.0138** | **0.0012** | **0.371** | **0.0138** | **0.0012** | **0.371** |
| Standard Deviation | 0.0066 | 0.0011 | 0.3107 | 0.0066 | 0.0011 | 0.3107 |

| | *pess* | | | *glai* | | |
|---|---|---|---|---|---|---|
| Outcome Feature | Qini | Top 20% Qini | Qini R-squared | Qini | Top 20% Qini | Qini R-squared |
| 50% Sample 1 | 0.0243 | 0.0021 | 0.803 | 0.0674 | 0.0077 | 0.864 |
| 50% Sample 2 | 0.0177 | 0.001 | 0.374 | 0.0588 | 0.0059 | 0.806 |
| 50% Sample 3 | 0.0125 | 0.0019 | 0.471 | 0.0656 | 0.0067 | 0.851 |
| 50% Sample 4 | 0.0081 | 0.0023 | -0.046 | 0.0571 | 0.0065 | 0.94 |
| 50% Sample 5 | 0.0278 | 0.0016 | 0.499 | 0.0698 | 0.008 | 0.908 |
| **Full Sample (100%)** | **0.0113** | **0.0020** | **0.814** | **0.0652** | **0.0084** | **0.835** |
| Standard Deviation | 0.0073 | 0.0005 | 0.2794 | 0.0066 | 0.0010 | 0.0519 |

**Figure 8**
Characteristics distribution of the patients from the synthetic dataset; the top 20 percent segment in which patients will benefit most from the treatment, and the bottom 20 percent segments in which patients will benefit less from the treatment



## 5. Discussion

Models to recruit patients for a specific treatment that are most beneficial to them have great potential to deliver the most effective and beneficial treatments to improve the health of patients. In this section, we will cover the approach and the limitations of this study which is followed by steps to overcome the limitations and lastly the results. We ran this study on evaluating treatment response uplift modeling to endorse the take-up of uplift modeling in clinical settings. As a result, we found that uplift modeling could be used to identify a subgroup of patients with heart failure who will benefit most from an HTM intervention based on its clinical characteristics. The main insight of the study is that although there is a large variability in terms of performance of some uplift modeling techniques, there is a promising technique, namely generalized Lai's approach (*glai*), performing consistently well and yielding robust models.

The objective of this study is to answer the research questions by experimentally evaluating uplift models on TEN-HMS dataset, comprising the data collected during a randomized clinical trial on assessing the effect of home telemonitoring in patients with heart failure. However, the dataset has its limitations. Due to the small size of it, 162 and 80 observations for the treatment and control groups, respectively, we could not reliably apply general model validation techniques, such as hold-out or k-fold cross-validation. Therefore, for the uplift model evaluation on the TEN-HMS dataset, we have conducted a self-test in which the test set to evaluate the uplift models is exactly the same as the training set to train the uplift models. Splitting the data into a training set and test set resulted in subsets, that were too small to obtain stable uplift models. Too few observations in the test set led to an uplift model evaluation of insufficient quality. As reported by Radcliffe and Surry (2011), for modeling binary outcomes, the product of the overall favorable response and the size of each population should be at least 500. For example, if the overall positive response rate is is 30%, there must be at least 1667 observations for each of the treated and the control group. In addition, many conventional validation methods specific for small datasets cannot be applied in our case. Notably, the leave-one-out cross-validation (LOOCV) is impractical for the evaluation of the uplift models with our dataset. As described in Section 3.2, the error of an uplift model cannot be computed from the predicted outcome of a single identity. Likewise, in the case of the leave-p-out cross-validation (LpO CV), the number of observation in the test set p has to be large enough to get reasonable results from uplift models.

To overcome the size limitations of the dataset, we generated 100 datasets having similar statistical characteristics with found in the TEN-HMS dataset, and combined them for our experiments. This lets us apply fivefold cross-validation to evaluate model performances and perform experiments to verify the stability of these models. Consequently, our results revealed that not all models yield similar performances and robustness across different experiments, except for *glai* approach, which in all experiments yields the highest uplift evaluation metrics and highest stability with lowest standard deviations in between those metrics. Contrary to the findings of Devriendt et al. (2018), the use of *glai* method yields the highest and the most stable results during our experiments. They have reported that they could not achieve stable and good results with this method on their experiments. The stability and the high performance of *glai* in our case might be due to the higher positive response rates than that of reported by Devriendt et

al. However, this hypothesis must be tested in future research with generated dataset samples with varying positive responds rates in their treatment and control groups.

The two-model approach (*tma*) was used as the baseline model during all experiments, instead of traditional classifier methods. In our case, it is theoretically wrong to use traditional classifier models as a baseline is because these models do not take into account what would happen if the treatment was not implemented. These models only predict the probability of responding rather than the increase in the probability of responding based on intervention.

Although accepted in academic literature, Qini metrics suffer from some limitations due to their dependencies on characteristics of the applications. They are not normalized measures; therefore, Qini values of uplift models obtained on different datasets for different treatment cannot be compared across datasets.

The results from the synthetic dataset are similar in quality as to those from TEN-HMS dataset. However, the performances are lower for the synthetic dataset. One of the main reason for this might be due to the testing of the model performances on unseen test sets, whereas, for TEN-HMS dataset, testing of the models is performed on the already seen dataset. This most probably led to overfitting of the models on TEN-HMS dataset and have an inflation impact on the performance.

With both datasets, we were able to map the characteristics distribution of the patients who will benefit most from the treatment by using generalized Lai's approach. It appears that the patient's characteristics as NYHA (New York Heart Association), BMI (Body Mass Index), a number of prior hospitalizations, and ejection fraction levels can be used together in the recruiting process of patients with heart failure for an HTM treatment. For example, for TEN-HMS dataset, it can be concluded that patients with NYHA level of two (mild symptoms with mild shortness of breath and/or angina, and slight limitation during ordinary activity), having BMI around 22 (healthful weight), less than 5 prior hospitalizations, together with the ejection fraction level of around 20 (pumping ability of the heart is severely below normal) can be recruited for an HTM treatment and they will most likely get the highest benefits from the treatment.

Overall, the results demonstrated that *glai* approach is a stable model, which can be recommended for datasets with varying sizes. However, for each application of uplift modeling, there are some important factors to consider. First, there should be a randomly selected control group in the dataset. Second, the treatment and the control groups from the dataset must have reasonable outcome observations. Although not supported by empirical or theoretical evidence, Kane et al. (2014) point out that a high "signal-to-noise ratio", the response rate difference between treatment and control groups over that of the control group, is important. Last, dataset size must be sufficiently large considering the stratified splitting for model evaluation and having a sufficient number of observations in both treatment and control groups.

## 6. Conclusion

Uplift modeling has tremendous potential in the health care field, especially in identifying people with chronic or life-threatening diseases who can be recruited for specific treatments where they will receive optimal benefits and health gains. In this study, we presented the applications of four different uplift modeling techniques on a real-world dataset from a randomized clinical trial testing an intervention in a heart failure patient cohort and a synthetic dataset comprising an intervention and a control group in synthesized heart failure patient cohort. We demonstrated and evaluated the implementation of two-model, dummy treatment, pessimistic, and generalized Lai's approaches. We showed that generalized Lai's approach (*glai*) is the model of choice with consistently higher performance and stability, in comparison to the other three methods.

During the study, we have built the models according to literature (Kane 2014, Lo 2002, Shaar 2016). As conventional evaluation methods are not applicable to uplift models, we have used specific evaluation methods for uplift modeling. The overall model evaluation approach that we used can be summarized as follows. We sort the observations from treatment and control groups in descending order of their predicted uplift scores, separately. Then, we divide them into different segments of equal size. Consequently, we take the pairwise differences of the uplift averages per segments to obtain an idea on the effectiveness of the treatment for each segment. For a more precise evaluation, Qini measure together with Qini curve was used, which depict the overall cumulative uplift due to the treatment effect on the certain fraction of population that ranked by the uplift model.

Results of the experiments, as presented in this study, highlight several conclusions. We can apply the uplift modeling methods on the real-world clinical trial dataset to stratify patients with heart failure into an HTM treatment with respect to patient characteristics at baseline. In particular, the implementation of the *glai* approach demonstrated that recruitment of the patients according to their NYHA (New York Heart Association), BMI (Body Mass Index), and a number of prior hospitalizations can be used together. For TEN-HMS dataset, it can be concluded that patients with NYHA level of two, having BMI around 22 together with less than 5 prior hospitalizations should be recruited for an HTM treatment for the highest benefit.

We have performed performance stability tests of the uplift methods in our generated dataset. In this regard, we took five different random samples having half of the observations of our generated dataset and compared the results with those from the entire dataset. For the stability experiments, the *glai* approach yields the highest performance metrics while showing less variance in the average Qini and Qini Top 20% measures per experiment.

To our knowledge, this is the first study, showing the application of uplift modeling methods to select the optimal treatment for patients with heart failure. Our study can be used as a guide to applying uplift models on datasets from clinical cohort studies. However, for each application of uplift modeling, there are some important factors to consider. The treatment and the control groups from the dataset must have reasonable outcome observations. Although not supported by empirical or theoretical evidence, Kane et al. (2014) point out that a high "signal-to-noise ratio" is important which is response rate difference between treatment and control groups over that of the control group. Dataset size must be sufficiently large considering the stratified splitting for model evaluation and having a sufficient number of observations in both treatment and control

groups. As Radcliffe and Surry (2011) reported, for modeling binary outcomes, the product of the overall favorable response and the size of each population should be at least 500. For example, if the overall favorable response is 0.1%, this means that both the treated and the control group need to be at least 500,000.

Finally, further research can be implemented to check the behavior of the uplift modeling approaches in multiple treatment setups. For example, we can make experiments with varying signal-to-noise, treatment-to-control groups' size, and respond observations ratios in order to have a better idea on the behaviors of those approaches. We can as well build models to predict continuous target variables. In addition, we can build further uplift models together with ensemble methods and neural networks to compare the performances on our generated datasets.

**References**

Austin, P. C., Mamdani, M. M., Juurlink, D. N., & Hux, J. E. (2006). Testing multiple statistical hypotheses resulted in spurious associations: a study of astrological signs and health. *Journal of Clinical Epidemiology*, *59*(9), 964–969.

Berezin, A. E., Kremzer, A. A., Martovitskaya, Y. V., Samura, T. A., Berezina, T. A., Zulli, A., … Kruzliak, P. (2015). The utility of biomarker risk prediction score in patients with chronic heart failure. *International Journal of Clinical and Experimental Medicine*, *8*(10), 18255–18264.

Biswas, A., Parikh, C. R., Feldman, H. I., Garg, A. X., Latham, S., Lin, H., … Wilson, F. P. (2018). Identification of Patients Expected to Benefit from Electronic Alerts for Acute Kidney Injury. *Clinical Journal of the American Society of Nephrology*, *13*(6), 842–849.

Chang, V., O'Donnell, B., Bruce, W. J., Maduekwe, U., Drescher, M., Mendez, B. M., … Patel, P. A. (2019). Predicting the Ideal Patient for Ambulatory Cleft Lip Repair. *The Cleft Palate-Craniofacial Journal*, *56*(3), 293–297.

Cleland, J. G. F., Louis, A. A., Rigby, A. S., Janssens, U., & Balk, A. H. M. M. (2005). Noninvasive Home Telemonitoring for Patients With Heart Failure at High Risk of Recurrent Admission and Death. *Journal of the American College of Cardiology*, *45*(10), 1654–1664.

Coussement, K., Lessmann, S., & Verstraeten, G. (2017). A comparative analysis of data preparation algorithms for customer churn prediction: A case study in the telecommunication industry. *Decision Support Systems*, *95*, 27–36.

Deaton, A., & Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine*, *210*, 2–21.

Devriendt, F., Moldovan, D., & Verbeke, W. (2018). A Literature Survey and Experimental Evaluation of the State-of-the-Art in Uplift Modeling: A Stepping Stone Toward the Development of Prescriptive Analytics. *Big Data*, *6*(1), 13–41.

García, S., Luengo, J., & Herrera, F. (2015). Data Preprocessing in Data Mining. *Intelligent Systems Reference Library*.

Guelman, L., Guillén, M., & Pérez-Marín, A. M. (2014). A survey of personalized treatment models for pricing strategies in insurance. *Insurance: Mathematics and Economics*, *58*(1), 68–76.

Holland, P. W. (1986). Statistics and Causal Inference. *Journal of the American Statistical Association*, *81*(396), 945–960.

Kane, K., Lo, V. S. Y., & Zheng, J. (2014). Mining for the truly responsive customers and prospects using true-lift modeling: Comparison of new and existing methods. *Journal of Marketing Analytics*, *2*(4), 218–238.

Kent, D. M., & Hayward, R. A. (2007). Limitations of applying summary results of clinical trials to individual patients: The need for risk stratification. *Journal of the American Medical Association*.

Koehler, F., Koehler, K., Deckwart, O., Prescher, S., Wegscheider, K., Kirwan, B. A., … Stangl, K. (2018). Efficacy of telemedical interventional management in patients with heart failure (TIM-HF2): a randomised, controlled, parallel-group, unmasked trial. *The Lancet*, *392*(10152), 1047–1057.

Kornegay, C., & Segal, J. B. (2013). Selection of Data Sources. *Developing a Protocol for Observational Comparative Effectiveness Research A User's Guide*, 109–124. Retrieved from

https://www.ncbi.nlm.nih.gov/books/NBK126190/pdf/Bookshelf_NBK126190.pdf
Kuhn, M. (2017). caret Package: Classification and Regression Training. *Https://Cran.R-Project.Org/*, 1–205.
Lai, L. Y.-T. (2006). Influential marketing: a new direct marketing strategy addressing the existence ofvoluntary buyers.
Lerman, R. I., & Yitzhaki, S. (1984). A note on the calculation and interpretation of the Gini index. *Economics Letters*, *15*(3–4), 363–368. https://doi.org/10.1016/0165-1765(84)90126-5
Lo, V.S.Y., & Lo, V. S. Y. (2002). The true lift model: a novel data mining approach to response modeling in database marketing. *ACM SIGKDD Explorations Newsletter*.
Lo, Victor S.Y., & Pachamanova, D. A. (2015). From predictive uplift modeling to prescriptive uplift analytics: A practical approach to treatment optimization while accounting for estimation risk. *Journal of Marketing Analytics*, *3*(2), 79–95.
Marinakos, G., & Daskalaki, S. (2017). Imbalanced customer classification for bank direct marketing. *Journal of Marketing Analytics*, *5*(1), 14–30.
Meinert, C. L. (2009). *Clinical Trials: Design, Conduct and Analysis*. *Clinical Trials: Design, Conduct and Analysis*.
Nowok, B., Raab, G. M., & Dibben, C. (2016). synthpop : Bespoke Creation of Synthetic Data in R. *Journal of Statistical Software*, *74*(11).
R Development Core Team. (2016). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*.
Radcliffe, N. J. (2007). Using control groups to target on predicted lift: Building and assessing uplift model. *Direct Marketing Analytics Journal*, (3), 14–21.
Radcliffe, N., & Surry, P. (2011). Real-world uplift modelling with significance-based uplift trees. *White Paper TR-2011-1, Stochastic Solutions*, (section 6), 1–33. Retrieved from http://www.stochasticsolutions.com/pdf/sig-based-up-trees.pdf
Rudaś, K., & Jaroszewicz, S. (2018). Linear regression for uplift modeling. *Data Mining and Knowledge Discovery*, *32*(5), 1275–1305.
Rzepakowski, P., & Jaroszewicz, S. (2010). Decision trees for uplift modeling. In *Proceedings - IEEE International Conference on Data Mining, ICDM*.
Rzepakowski, P., & Jaroszewicz, S. (2012a). Decision trees for uplift modeling with single and multiple treatments. *Knowledge and Information Systems*, *32*(2), 303–327.
Rzepakowski, P., & Jaroszewicz, S. (2012b). Uplift modeling in direct marketing. *Journal of Telecommunications and Information Technology*, *2012*(2), 43–50.
Sc, I., Ra, C., Dierckx, R., & Jgf, C. (2015). Structured telephone support or non-invasive telemonitoring for patients with heart failure ( Review ), (10). https://doi.org/10.1002/14651858.CD007228.pub3.www.cochranelibrary.com
Shaar, A., & Segard, O. (2016). Pessimistic Uplift Modeling. https://doi.org/10.475/123
Sibbald, B., & Roland, M. (1998). Understanding controlled trials: Why are randomised controlled trials important? *BMJ*, *316*(7126), 201–201.
Sołtys, M., Jaroszewicz, S., & Rzepakowski, P. (2015). Ensemble methods for uplift modeling. *Data Mining and Knowledge Discovery*, *29*(6), 1531–1559.
Yun, J. E., Park, J. E., Park, H. Y., Lee, H. Y., & Park, D. A. (2018). *Comparative Effectiveness of Telemonitoring Versus Usual Care for Heart Failure: A Systematic Review and Meta-analysis*.

*Journal of Cardiac Failure* (Vol. 24). Elsevier Inc.

Zaniewicz, L., & Jaroszewicz, S. (2013). Support Vector Machines for Uplift Modeling. In *2013 IEEE 13th International Conference on Data Mining Workshops* (pp. 131–138). IEEE.