HHS Public Access

Author manuscript

Inform Med Unlocked. Author manuscript; available in PMC 2023 November 20.

Published in final edited form as:

Inform Med Unlocked. 2023 ; 42: . doi:10.1016/j.imu.2023.101362.

Personalized statin treatment plan using counterfactual approach with multi-objective optimization over benefits and risks

Yue Liang[a,e,*], **Pui Ying Yew**[a,e], **Matt Loth**[a,b,e], **Terrence J. Adam**[a,c], **Julian Wolfson**[d], **Peter J. Tonellato**[f], **Chin-Lin Chi**[a,b,e]

[a]Institute for Health Informatics, University of Minnesota, Minneapolis, MN, 55455, USA

[b]School of Nursing, University of Minnesota, Minneapolis, MN, 55455, USA

[c]Department of Pharmaceutical Care & Health Systems, University of Minnesota, Minneapolis, MN, 55455, USA

[d]Division of Biostatistics, University of Minnesota, Minneapolis, MN, 55455, USA

[e]Optum Labs Visiting Fellow, Eden Prairie, MN, 55344, USA

[f]Department of Health Management and Informatics, University of Missouri School of Medicine, Missouri, 65212, USA

Abstract

Background: Statins are a class of drugs that lower cholesterol levels in the blood by inhibiting an enzyme called 3-hydroxy-3-methyl-glutaryl-coenzyme A (HMG-CoA) reductase. High cholesterol levels can lead to plaque buildup in the arteries, which can cause Atherosclerotic Cardiovascular Disease(ASCVD). Statins can reduce the risk of ASCVD events by about 25–35% but they might be associated with symptoms such as muscle pain, liver damage, or diabetes. As a result, this leads to a strong reason to discontinue statin therapy, which increases the risk of cardiovascular events and mortality and becomes a public-health problem.

To solve this problem, in the previous work, we proposed a framework to produce a proactive strategy, called a personalized statin treatment plan (PSTP) to minimize the risks of statin-associated symptoms and therapy discontinuation when prescribing statin. In our previous PSTP framework, three limitations remain, and they can influence PSTP usability: (1) Not taking the counterfactual predictions and confounding bias into account. (2) The balance between multiple drug-prescribing objectives (especially trade-off objectives), such as tradeoff between benefits and risks. (3) Evaluating PSTP in retrospective data.

[*]Corresponding author. Institute for Health Informatics, University of Minnesota, Minneapolis, MN, 55455, USA. liang775@umn.edu (Y. Liang).

Declaration of competing interest

No conflict of interest.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.imu.2023.101362.

Objectives: This manuscript aimed to provide solutions for the three abovementioned problems to improve PSTP robustness to produce a proactive strategy for statin prescription that can maximize the benefits (low-density lipoprotein cholesterol (LDL-C) reduction) and minimize risks (statin-associated symptoms and therapy discontinuation) at the same time.

Methods: We applied overlapping weighting counterfactual survival risk prediction (CP), multiple objective optimization (MOO), and clinical trial simulation (CTS) which consists of Random Arms, Clinical Guideline arms, PSTP Arms, and Practical Arms to improve the PSTP framework and usability.

Results: In addition to highly balanced covariates, in the CTS, the revised PSTP showed improvements in lowering the SAS risks overall compared to other arms across all time points by at most 7.5% to at least 1.0% (Fig. 8(a)). It also has the better flexibility of identifying the optimal Statin across all time points within one year.

Conclusion: We demonstrated feasibility of robust and trustworthy counterfactual survival risk prediction model. In CTS, we also demonstrated the PSTP with Pareto optimization can personalize optimal balance between Statin benefits and risks.

Keywords

Statin; Statin-associated-symptoms; Counterfactual prediction; Generalized propensity score; Generalized overlap weights; Clinical trial simulation

1. Introduction

Statins are a class of drugs that inhibit cholesterol synthesis by blocking hydroxymethylglutaryl coenzyme-A (HMG-CoA) reductase, an enzyme involved in cholesterol production. They are widely prescribed for patients with high cholesterol levels or at risk of cardiovascular disease, as they have been shown to reduce the incidence of myocardial infarction, stroke, and mortality [1]. However, statins have the potential to cause muscle pain, liver damage, renal failure, and rhabdomyolysis, collectively referred to as statin-associated symptoms (SAS). Some other lieratures pointed out some other specific statin related side effects including myositis [2]. While recent work has called into question the degree to which SAS are due to the medication versus nocebo effects [3], it remains true that concerns about SAS can reduce adherence and lead to discontinuation of statin therapy [1]. To minimize the impact of this problem, in our previous study [4], we proposed a proof-of-concept tool to produce a proactive strategy called Personalized Statin Treatment Plan (PSTP), which utilizes a machine-learning approach to personalize the treatment plan (specific statin agent and dosage) with the minimal SAS and therapy discontinuation risks.

However, due to the research scope, there are three important points that were not addressed in our previous PSTP model: (1) Not taking the confounding bias into account, (2) not having the balance between benefits and risks during the optimization procedure, such as the balance between the benefit of LDL-C reduction and risks of SAS and discontinuation, (3) not having the evaluation of the proactive prescribing strategy (a.k.a, PSTP) in retrospective data. These points are essential to evolving the PSTP on the way toward practical usage, and we aimed to address them in this article.

For the first point, the original PSTP method [4] discussed using a predictive model to estimate patient outcomes of a treatment plan, but did not address the potential confounding bias that might arise in observational data. The potential confounding bias may influence the outcome estimation, which is highly relevant to PSTP usability. This is commonly known as the counterfactual problem. Here, we resorted to the generalized propensity score and Overlap Weights as the solution to perform counterfactual predictions and adjust for confounding issues. More specifically, one makes a more robust estimation under multiple treatment setting and causal perspective by generating the pseudo population with more emphasis on clinical equipoise, which means that it focuses on subpopulation with larger probability to be assigned to all treatment options. For example, patients with larger uncertainty to be assigned to a specific type of statin are emphasized with larger weights than patients with smaller uncertatinty to be assigned to a specific statin treatment [5]. This gives us a more robust counterfactual prediction results under causal assumptions that patients that are in clinical equipoise have more credible outcomes for each other to borrow.

For the second point, the previous work only minimized SAS and discontinuation risks [4] to produce PSTP, which limits the PSTP usability. To improve usability, we also need to maximize statin benefits (aka. LDL-C reduction) where the lower LDL-C value, the better for patients [6]. In this manuscript, we also discussed the use of multi-objective optimization to optimize the balance between maximal LDL-reduction benefit and minimal SAS and discontinuation risks [7]. In our statin project, Multi-objective optimization refers to the optimization procedure involving 3 objective functions: $(f_t^{SAS}, f_t^{Disc}, f_t^{-LDL-C\ reduction})$.

For the third point, the gold standard approach to evaluate the usability of PSTP is to conduct a randomized clinical trial (RCT), but RCTs are expensive and time-consuming. Therefore, we proposed a clinical trial simulation (CTS) as an alternative option to evaluate the PSTP and give us baseline information about the potential improvement in a well-controlled setting. Briefly, we simulated the prospective setting by using counterfactual prediction and compared PSTP arm of treatment with three other arms of treatment. These three arms are the clinical guideline, clinical practice, and randomization. The patient outcomes improvement of the proactive strategy, PSTP, can then be readily comprehended through these comparisons.

In addition, PSTP usage is prospective and involves different time points. The PSTP is the optimal prescription that minimizes the probability of a SAS event in a year. Intead of doing intent-to-treat analysis as our previous study did, per-protocal approach using survival model is adopted as we assume that patients will be adhereing to the treatment for at least one year.

2. Materials and methods

The whole proactive strategy for selecting the optimal Statin treatment plan consists of three major steps: data extraction and preprocessing, model development and prediction, and model evaluation and validation. The first step identifies important variables and possible confounders related to SAS, discontinuation, and LDL-C reduction, by combining multiple data sources, including demographic data, claims data, and laboratory data, and diagnostic data. The second step produces the counterfactual prediction using a counterfactual

approach. The third step implements Muti-Objective Optimization (MOO) on the predicted outcomes to select the optimal treatment recommendations for each patient. The final step evaluates the PSTP treatment in CTS by comparing it to the random arm, clinical practice arm, and statin clinical guideline arm.

2.1. Data extraction and preprocessing

The data extraction procedure specifies the required data for step 2 – model development and prediction. We have selected variables based on clinical advice. The following specifies how we defined, extracted, and preprocessed for Statin treatment variables, outcome variables, and other baseline characteristics.

Definition of the treatment variable: Statins are a class of drugs with multiple combinations of Statin generic type and dosages. Based on LDL-C reduction intensity, it can also be categorized into low, moderate, and high intensity level for different amount of the Statin dosage. We summarized statin agents and dosages used in this study in Table 1.

Definition of the clinical endpoints/outcomes—In our Statin study, the three main clinical endpoints are SAS, discontinuation of the statin treatment, and reduction of LDL-C SAS is defined based on the diagnosis records using ICD-9/10 code, which mainly includes rhabdomyolysis, myopathies, renal events, liver events and poisoning events, etc. Discontinuation is defined as occurring when a patient has a gap in statin supply that last for more than 30 days. Patients who had only one prescription fill and never got refill after the first prescription are labeled as censored. LDL-C reduction is defined as the subtraction of 3 month average between pre-index and post-index LDL-C value, where index date of medication is defined as the earliest date of the prescription.

Definition of confounding variables or other baseline characteristics— Addressing confounding variables is crucial because failing to do so can lead to biased prediction. We used comorbidity flags (binary indicator) extracted from diagnosis table as the major confounding variables for our analysis. We included all existing baseline variables and comorbidity flags into the model to try our best to achieve the fulfillment of no unmeasured confounding assumption. The details of the comobidity flags are specified in the Supplementary Table 1 and Table 2.

The data used for the Statin users is from the OptumLabs® Data Warehouse. It has over 126.6 million of the claims data and 108 millions Electronic Health Records (EHR)data records. The data extraction contains adults' records who took Statin treatment in their medication table. Other related data tables contain the basic demographic information, patients' insurance claims data, electronic health records including diagnosis and comorbidities information, and laboratory results.

The study cohort included all Statin users from 2015 to 2022 with at least a 1 year clean period. The patients with at least 1 year clean period means that patients have Statin prescription for the first time with at least 1 year pre-index period without taking any statin medications. The study cohort was required to have continuous medical and pharmacy insurance coverage prior to their statin index date. The baseline period is defined as one year

prior to the Statin index date and one year after index date regarded as the follow-up period. Pre-index and post-index LDL-C value is also required to be present so that we can define the LDL-C reduction as one of the clinical outcomes after patients took Statin medication. The statin cohort has 179,413 number of patients.

In our data, patients who were prescribed any added/combination drugs were excluded. Any combination or added drugs with Statin might have interference in which the prediction of the SAS and discontinuation of the Statin drug can be inaccurate because we are not sure which treatment contributed to SAS or discontinuation when they happened. Patients who have had their medications paid by Medicare were also excluded from the study. Furthermore, only patients within age range from 20 to 75 were included in the study. Data phenotyping and preprocessing procedure is done based on clinical domain knowledge, including selecting possible related confounding variables from the data, data phenotyping procedure, data imputation feasibility, and potential existing bias. Variables with more than 80% of patients with missing values were excluded from the data. Data imputation is proceeded using the iterative k-nearest neighbor methods [8]. In addition, patients who took Lovastatin with moderate intensity and Pitavastatin with low intensity were excluded from data because of the low sample size of these two treatment plans. Lovastatin with moderate intensity only has 516 patients (0.28%) and Pitavastatin with low intensity only has 181 patients (0.1%). After removing those two statin treatment arms, the total cohort sample size became 178,716. Patients with no observed of both SAS event and time-to-event in days (2,483) are removed. The final cohort size for analytical data is 176,233. Total number of statin treatment options is 10 (Atorvastatin High intensity (AT_High), Atorvastatin Moderate intensity(AT_Moderate), Lovastatin Low intensity(LO_Low), Pitavastatin Moderate intensity(PI_Moderate), Pravastatin Low intensity(PR_Low), Pravastatin Moderate intensity (PR_Moderate), Rosuvastatin High intensity(RO_High), Rosuvastatin Moderate intensity(RO_Moderate), Simvastatin Low intensity(SI_Low), Simvastatin Moderate intensity(SI_Moderate)).

The data extraction and cohort selection workflow for the Statin cohort is summarized in Fig. 1.

2.2. Model development settings

To minimize the influence of confounding variables and optimize the balance between benefits and risks to improve PSTP usability, we applied a counterfactual prediction (CP) and multiple-objective optimization (MOO) in this project. More specifically, to develop the CP, we created a Generalized Propensity Score predictions through multi-class classification using Neural Networks. Then, we calculated the Overlap Weights (OW) based on the Generalized Propensity Score [5]. Finally, we trained outcome models using the OW as the sample weight for each treatment group. For MOO, we use the Technique for Order of Preference by Similarity to the Ideal Solution(TOPSIS) to minimize SAS and discontinuation, and maximize LDL-C reduction to generate the PSTP.

2.2.1. Counterfactual Prediction(CP): esitmate the generalized propensity score (GPS)—

Generalized propensity score estimation is a method for estimating causal

effects of continuous treatments or multiple interventions based on observational data. The propensity score is defined as the conditional probability of receiving a certain of treatment options given a set of covariates [9]. In the meantime, Multilayer perceptron (MLP) is a type of artificial neural network that can be used to estimate generalized propensity scores by learning a nonlinear function that maps covariates to treatment propensity scores. MLP can handle complex and high-dimensional data and may provide more accurate and flexible estimates than other parametric models [10].

In our statin project, we used the MLP as the multi-class classification model to make generalized propensity score estimation. Each class stands for the statin treatment options. The sum of the predicted probabilities of each statin treatment equals to 1 for each data point (patient). The initial input layer contains 121 number of units, which is corresponding to the number of independent features of the data fed into the model. The first hidden layer contains 968 number of hidden units, which is 8 times of the number of initial input layer units. The second hidden layer has 484 number of hidden units, which is half number of the first hidden layer. We use 8 times as many input nodes because we want to ensure the model capacity is as large as possible accompanying with regularization method to prevent overfitting. Dropout regularization with 30% dropout rate and early stopping of the training procedure is carried out in our model at the best validation performance training epoch [11]. ReLU function is used as the activation function between layers. The final output layer consists of 10 units which are corresponding to the number of treatment options and we used cross entropy function as our loss function. After the training procedure is done through backpropagation on the entire dataset, it is then used again to feed into the trained model and GPS is estimated using the softmax function.

The following equation shows the GPS estimation for each individual:

$$Pr^{i}(T = t \mid \boldsymbol{X}) = e_t^i(\boldsymbol{X}),\ s.t.\ \sum_{t=1}^{T} e_t^i = 1$$

$T = t$ denotes the treatment labels for the MLP model. $\boldsymbol{X}$ denotes the independent features fed into the MLP. e_t^i stands for the classification probability of getting treatment t given the independent features matrix $\boldsymbol{X}$ for each individual i. It essentially is a multi-class classification problem with all predicted treatment probability summing up to 1.

The predicted GPS is a matrix with dimension $N \times T$ where N stands for the number of data records(patients), and T stands for the number of treatment options. Recall that in our statin project, the number of treatment options is 10.

Algorithm 1.

CF prediction using Generalized Overlap Weights

Require: $t \in \mathbb{T}$
$Train\boldsymbol{e}_t(\boldsymbol{X}) := Pr(T = t \mid \boldsymbol{X}), t \in \mathbb{T}$

Get Generalized Propensity Score prediction $\widehat{\boldsymbol{e}}_t(\boldsymbol{X}) \in \mathbb{R}^{N \times \lvert\mathbb{T}\rvert}$	
Generate Generalized Overlap Weights:	
$$\boldsymbol{GOW} := \frac{\left(\sum_{t=1}^{T} \frac{1}{\hat{e}_t(\boldsymbol{X})}\right)^{-1}}{\hat{e}(\boldsymbol{X})}, t \in \mathbb{T}, \boldsymbol{GOW} \in \mathbb{R}^{\mathrm{N} \times \lvert\mathbb{T}\rvert}$$	
for $\boldsymbol{t} = \boldsymbol{1}, \boldsymbol{2}, \boldsymbol{3}, \ldots, \boldsymbol{T}$ ***do***	
$$f_t := \mathbb{E}(< \boldsymbol{Y}_t, \boldsymbol{C}_t > \mid \boldsymbol{X}_t; \boldsymbol{GOW}_t), t \in \mathbb{T}$$	
CF prediction: $\widehat{\boldsymbol{Y}}^{(\boldsymbol{t})} = \widehat{f}_t(\boldsymbol{X})$	
end for	
Notations for Algorithm 1:	
$\mathbb{T}$:	Treatment set
$\lvert\mathbb{T}\rvert$:	Cardinality of the treatment set. In our statin project, it equals = 10
$\boldsymbol{T}$:	Random variable: treatment
$\Pr(T = t \mid \boldsymbol{X})$:	Probability of randome variable $\boldsymbol{T}$ is equal to the specific treatment t, given independent features in matrix $\boldsymbol{X}$
$\boldsymbol{X}$	Independent features matrix
$\boldsymbol{e}_t(\boldsymbol{X})$	Probability matrix $\in \mathbb{R}^{\mathrm{N} \times \lvert\mathbb{T}\rvert}$
$\widehat{\boldsymbol{e}}_t(\boldsymbol{X})$	Predicted/estimated probability matrix $\in \mathbb{R}^{\mathrm{N} \times \lvert\mathbb{T}\rvert}$ by the model
$\boldsymbol{GOW}$	Calculated Generalized OW matrix $\in \mathbb{R}^{\mathrm{N} \times \lvert\mathbb{T}\rvert}$ using the estimated probability matrix $\widehat{\boldsymbol{e}}_t(\boldsymbol{X})$
$\boldsymbol{X}_t$	Independent features of submatrix of $\boldsymbol{X}$ for treatment cohort $T = t$
$\boldsymbol{GOW}_t$	Sub-matrix of the generalized OW for the treatment cohort $T = t$
$\boldsymbol{Y}_t$	Sub-vector of the outcome variable $\boldsymbol{Y}$ for treatment cohort $T = t$
$\boldsymbol{C}_t$	Sub-vector of the censoring indicator $\boldsymbol{C}$ for treatment cohort $T = t$
$\widehat{\boldsymbol{Y}}^{(\boldsymbol{t})}$	Esitmation of the counterfactual outcome for the whole cohort for treatment t

2.2.2. CP: calculate the generalized overlap weights (GOW)—After estimating the generalized propensity score using MLP from the above procedure, this step aims to calculate the Generalized Overlap Weights (GOW) based on the predicted generalized propensity scores [5, 12]. We chose GOW for three reasons: (1) It can incorporate more than 3 treatments without cutting off any data points. (2) It emphasizes on the clinical equipoise where patients with larger uncertainty of being assigned to any treatments will be up-weighted, and patients with a certain probability of treatment assignment will be down-weighted. (3) It has a smaller variance due to its mathematical property.

The following formula defines the calculation of the Generalized Overlap Weights for each individual i:

$$\mathbf{GOW^i}: = \frac{\left(\sum_{t=1}^{T} \frac{1}{e_t^i}\right)^{-1}}{e^i}, t \in \mathbb{T}$$

where e^i, GOW^i are vectors with length of $|\mathbb{T}|$.

After the calculation, the generalized OW have the same dimension as the GPS ($N \times |\mathbb{T}|$) where each data point (patient) has their own weight for each statin treatment with dimension with length $|\mathbb{T}|$.

2.2.3. CP: training and applying outcome models using the GOW as the sample weight for each treatment group—The outcome model can be trained using supervised machine learning model with outcome vector $\boldsymbol{Y}$ and independent features matrix $\boldsymbol{X}$. The generalized overlap weights are used as sample weights during training. In our statin study, we are generating counterfactual predictions of SAS, discontinuation risks, and LDL-C reduction for each patient if they continuously used different Statin treatment for a year. This is helpful to doctors' prescription procedure because they know that if a patient is considering ceasing from taking statin treatment, the benefit (or side effect) will likely stop. In order to understand what would be the effect after patients received assigned treatment, we adopted per-protocol analysis where patients were assumed to be adhering to their treatment for the entire follow-up period [13]. Here we chose to use survival model to follow the per-protocol analysis to incorporate time-to-event information in predicting the survival probability of SAS and discontinuation in one year, while a linear model with regularization was used to make counterfactual predictions for LDL-C reduction since LDL-C reduction is a continuous variable and it would be quite difficult to proceed survival model for continuous outcome.

The model for SAS and discontinuation can be specified as follows:

$$f_t = \mathbb{E}(\langle \boldsymbol{Y}_t, \boldsymbol{C}_t > | \boldsymbol{X}_t, \boldsymbol{GOW}_t), t \in \mathbb{T}$$

where model f_t is trained with the observed outcome, and Y_t is defined as the time to event in days. $\boldsymbol{X}_t$, $\boldsymbol{GOW}_t$ are corresponding sub-matrix of the data $\boldsymbol{X}$ and matrix $\boldsymbol{GOW}$. Variable C_t is defined as the censoring indicator for each treatment cohort $T = t$. Here the training data for f_t is the corresponding statin treatment cohort data. The corresponding $\boldsymbol{GOW}_t$ for each cohort will be normalized and fed into the model as the sample weight during the training process. Patients who are more indeterminate to be assigned a specific treatment will have higher weight, while patients who have smaller uncertainty to be assigned to a specific statin treatment will receive smaller weight. Therefore, there will be $|\mathbb{T}| = 10$ number of models f_t. For each model f_t, we use it to make the counterfactual prediction for the whole data X and predict the counterfactual outcome for treatment t (e.g., Atorvastatin with low-intensity level). Therefore, there will be a total $|\mathbb{T}|$ sets of predictions. Then, we use the trained model

to predict survival function and it will give us the survival probability $P(Y_t > y)$, and then we use $1 - P(Y_t > y) = P(Y_t < y)$ to retrieve the corresponding cumulative SAS event risks and discontinuation event risks within the time period between statin index date and specific time point y.

The model LDL-C percentage reduction is specified as the following:

$$f_t = \mathbb{E}(\boldsymbol{Y}_t \mid \boldsymbol{X}_t, \boldsymbol{GOW}_t), t \in \mathbb{T}$$

where f_t is trained with the observed outcome LDL-C percentage reduction as a continuous variable. Similarly, the corresponding GOW for each cohort is normalized and incorporated into the model as sample weight during the modeling fitting procedure.

The final counterfactual predictions utilized the independent variables X_t again when building up the outcome model for each treatment cohort. We utilized Random Survival Forest and Cox Proportional hazard model to develop survival models for SAS and discontinuation, and used Linear regression with L1 regularization (LASSO) to develop model for LDL-C reduction [14].The LASSO model was built with 5-folds cross validation.

The entire counterfactual modeling procedure (as shown above) is summarized in Algorithm 1.

2.3. Multi-objective optimization using the predicted counterfactual outcomes to choose the PSTP

We fed patients' data into the model and predicted the outcome at every 30 days from the 30th day up to the 360th day (with every 30 days increment) risks of SAS and discontinuation. The predicted survival function gives us the survival probability $Pr(Y_t > y)$ at each time point y ($y = 30$ days, 60 days, etc.). We then use $1 - Pr(Y_t > y) = Pr(Y_t \leq y)$, which means the probability of having the event within y time period. We also fed the patients into the counterfactual model for LDL-C reduction and predicted the counterfactual prediction of percentage LDL-C reduction for the simulated patient. Simulated patients' optimization result is showed in result section. We use the multi-objective optimization to determine PSTP for each individual. Specifically, we considerd three objectives: 1) minimizing the SAS risks, 2) minimizing the discontinuation risks, and 3): maximizing LDL-C reduction. We used the Technique for Order of Preference by Similarity to the Ideal Solution (TOPSIS) [15] where it utilizes the Euclidean distances to select the best solution from the pareto front optimal solutions with 3-dimensional spaces with x, y, z-axis corresponding to predicted SAS risks, discontinuation risks, and LDL-C percentage reduction respectively. We want to find the t^* that it can minimize SAS and discontinuation, while maximize LDL-C reduction (equivalent to the goal of minimizing (– LDL Reduction)). It can be specified as the following optimization problem:

$$\begin{aligned} &min_{(t \in T)} : \left(f_t^{SAS}, f_t^{Disc}, f_t^{-LDL-C\ reduction}\right) \\ &subject\ to : t \in T \end{aligned}$$

where f_t are the counterfactual predictions of SAS, Discontinuation risks and LDL-C reduction. With this step, we aimed to choose the best statin treatment plan (i.e., PSTP) by choosing the best Pareto Frontier solution in the 3-dimensional space of predicted SAS, Discontinuation and negative LDL-C reduction [7].

2.4. Model validation and evaluation

2.4.1. Balancing check and overlap check—In order to ensure better validity and robustness of the counterfactual prediction result under no unmeasured confounding assumption, we conducted the balancing and overlap check using the predicted propensity score to avoid potential confounding bias in retrospective study through GOW. Two evaluation metrics called Population Standard Mean Difference (PSD) and Absolute Standard Mean Difference (ASD) for weighted and unweighted covariates can be used to see if GOW can balance the confounding variables and eliminating possible confounding biases [16]. Overlap check is proceeded by plotting out the weighted and unweighted covariate distribution to see if weighted one achieves better overlap for all treatment groups. Weights are normalized to redistribute the covariate distribution so that each treatment cohort has better overlap with each other to achieve more robust mutual counterfactual prediction for each cohort, which gives us stronger confidence to conclude that patients in different treatment groups are exchangeable given covariates.

2.4.2. Clinical trial simulation—PSTP is a proactive strategy identified from retrospective data aiming for maximal benefit and minimal risk for an individual. We propose using clinical trial simulation (CTS) as the evaluation for this proactive strategy.

Briefly, we compare PSTP arm with the other three treatment arms. The first treatment arm is the random Arm, where each patient is assigned to a random treatment option. The second Arm is Clinical Guideline arm, where patients are assigned to treatment options under the 2013 ACC/AHA Clinical Guidelines [17]. The third Arm is the Clinical Practice arm, which is just the observed treatment options in the data. The last Arm is our PSTP Arm. We compared SAS risk, Discontinuation risk, and LDL-C reduction averaged across every individual in that arm. Every patient was assigned to every CTS arms. As a result, CTS can only give baseline information about the outcome improvement and evaluation in a well controlled environment.

3. Results

3.1. Data result

Table 2 presents patient baseline characteristics across different statin intensity levels and agents. The full table can be found in the supplementary material. The columns stands for the used statin agent name across our database and rows are the related clinical varaibles used for modeling. In Table 2, moderate intensity accounted for the larger proportion of the total statin users. The pre-index LDL-C level showed a higher value for High and Moderate cohorts, while the low-intensity cohort's pre-index LDL-C level was relatively lower than the other two cohorts. This is intuitive that doctors might make the prescription based on patients' pre-LDL-C lab value. Table 2 also shows Atorvastatin, Pravastatin, Rosuvastatin,

and Simvastatin are more frequently prescribed comparing to the other two types of statins. Several covariates showed differences across different statin groups. Interestingly, Pitavastatin group tends to have a larger proportion of the patients who had or were having lipid regulating types of the prescriptions.

3.2. Balancing check

Fig. 2 shows the PSD and ASD calculation for each covariate in our data. PSD calculates the standardized differences between each treatment cohort covariate mean and the population means across all treatment arms. The red lines show the weighted result and the blue line shows the unweighted result. ASD calculates a pairwised standard difference between covariates within every single exsiting pairs of covariates. Similarly, Red lines give us the weighted result, and the blue lines stand for unweighted result. Fig. 2showed how OW performed to balance out the confounding covariates in our statin data. The closer to 0 of the red lines, the better balancing result we are achieving using GOW among all statin treatment arms. Even though GOW did not make the PSD and ASD to be 0, it shows a considerable improvement in balancing out the variables in the dataset.

3.3. Overlap check

In Fig. 3, we presented several important covariate overlap check for those 10 Statin treatment plans. Here we use a typical confounding variable age as an example. The weighted distribution (Fig. 3(a)) has better and closer overlap for statin PI_moderate treatment option for the age variable comparing to unweighted distribution (Fig. 3(b)). Another two overlap check for Charlson Quan comorbidity score and pre-LDL-C values are shown in supplementary table.

3.4. Counterfactual prediction

Figs. 4 and 5 show examples of ten individualized counterfactual survival plots for SAS and discontinuation for four treatment plans (AT_High, AT_Moderate, SI_Low, and RO_Moderate). These plots were produced by the SAS and discontinuation counterfactual survival models. We use the trained model to predict survival function and it will give us the survival probability $P(Y_t > y)$, and then we use $1 - P(Y_t > y) = P(Y_t < y)$ to retrieve the corresponding cumulative SAS event risks and discontinuation event risks within the time period between statin index date and specific time point y. More surivival curves plots for all statin treatments are provided in supplement.

3.5. Optimization showcase

Here, we demonstrated our pipeline to select the personalized statin treatment plan (PSTP) for 10 simulated patients using multi-objective optimization when minimizing SAS risks, minimizing discontinuation risks, and maximizing LDL-C reduction at 30–360 days. In the illustration shown in Fig. 6, Atoravatstatin with High intensity level, Atorvastatin with Moderate intensity level, Pitavastatin with Moderate intensity level, Rosuvastatin with High intensity level, and Simivastatin with Moderate intensity level (Fig. 6(a) in purple) were selected at 30 days as the Pareto front optimal treatment after thorough consideration of SAS risks, discontinuation risks, and LDL-C reduction for patient, as shown in purple color.

Among all Pareto front optimal treatment options, TOPSIS selected one of the best from those treatment plans in purple. At different time points, the selected optimal treatments can be different. There are fewer pareto optimal treatments when increase the number of days since statin index date.

3.6. Clinical trial simulation

Fig. 7 represents the summary of the counts each CTS arm for all statin treatment options. We chose 180 days, 210 days, and 360 days survival risks as examples when selecting out the optimal treatment. Pitavastatin with moderate intensity (Fig. 8) was the most recommended by the optimization based on 180 days survival risks. For the survival risks within 1 year, Atorvastatin with high intensity, Atorvastatin with moderate intensity, and Rosuvastatin with moderate intensity were the most recommended when considering all 3 prediction targets. The Rosuvastatin with high intensity was the most recommended by optimization for 210 days of the survival risks for both SAS and Discontinuation risks.

Fig. 8depicts the mean value of predicted survival risks over time for each treatment arms for our cohort data. The PSTP Arm calculated using multi-objective optimization performed the best in SAS. The elbow in Fig. 8(a) in 180 days and 210 days showed that SAS prediction increases rapidly after 180 days of the study period. For the discontinuation risks, the PSTP showed similar risks compared to other arms. When it comes to LDL-C reduction (Fig. 8(c)), the value is higher when patients can stay on the treatment for a longer time. The reason why Fig. 8(c) only has the horizontal line for Random, Guideline, and Practice arms is that the outcome variable LDL-C reduction is calculated as the subtraction of the pre-index average LDL-C value and post-index average LDL-C value, which is static over time. LDL-C are recorded in laboratory data and the calculation of the LDL-C reduction is static. Our PSTP is determined by MOO considering SAS, Discontinuation and LDL-C reduction, so the dynamic SAS and discontinuation survival risks gives us dynamic PSTP arm across different time points, which is corresponding to different LDL-C reduction in Fig. 8(c). In general, we see PSTP has a lower SAS risk and similar discontinuation risk compared to all other arms. The PSTP selected in later period has a better LDL-C reduction as shown in Fig. 8(c).

4. Discussion

The updated PSTP framework in this article fully addressed the three major issues in our previous study. Counterfactual prediction and the potential confounding issue were addressed through incorporating GPS and GOW into our modeling procedure. In addition, we used LDL-C percentage reduction as an example of a benefit and demonstrates how we maximize benefits and minimize risks in the 3D MOO process. Finally, we used clinical trial simulations to evaluate the proactive strategy, PSTP.

For counterfactual prediction, we used the GOW proposed by Li et al. [5] to generalize the multiple treatment comparisons. This inspired us to use OW to build up 10 statin outcome models for each statin treatment. The original Li's article showed very balanced covariates across different treatment arms, while our GOW for Statins treatment arms does not make the covariates to be balanced to 0 in Fig. 2. Larger number of treatment arms in causal

related analysis makes the propensity score to be smaller to achieve the clinical equipoise, in which case it would be easier to generate extreme weights and harder to achieve the fully balance for covariates across all treatment arms.

Instead of doing intention-to-treat analysis, the individualized survival model provided us the flexibility of retrieving risks on the per-protocol basis, where the survival functions give us dynamic predictions for an individual's SAS and discontinuations risks on one year follow-up period. Rather than using SAS and discontinuation as dichotomous labels for model training, doctors or patients have the flexibility to choose which time point event risks they care about. The PSTP along all time points showed better SAS results while it didn't show a better result for discontinuation. The predicted SAS result has larger influence on deciding the PSTP optimal treatment options (Fig. 8). We could have different weights for those three factors(SAS, Discontinuation, LDL-C reduction) for deciding the PSTP in the future work. In addition, LDL-C reduction data is not abundant enough to do dynamic analysis across all time points. We can fit LDL-C reduction model for each time point to generate the dynamic plot like SAS and Discontinuation in the future work if data allows.

In our study, the Random survival forest was used for SAS and Discontinuation risk prediction, and the linear regression with L1 regularization was used for LDL-C reduction prediction. The extrapolation using linear models is critical to treatment cohorts especially when the sample size is small. Due to lack of data such as Lovastatin or Pitavastatin, the model is not able to fully learn the characteristics of patients who were prescribed for these two treatments. Linear models with extrapolation ensured that we could make individualized predictions [18].

By adding the LDL-C percentage reduction as another new objective for the modeling procedures, the updated PSTP fully addressed the balance between benefits and risks when making data-driven prescriptions. The individual 3D optimization shows an example of deciding statin treatment at different time points. In general, the longer time after Statin index date, the fewer optimal Statin options. This is because patients have higher probabilities of experiencing SAS and/or discontinuation when we have a sufficiently long time of using statin therapies. By taking advantage of EHR data and claims records, our counterfactual survival model framework, therefore, is able to identify the optimal treatment options at different time periods.

There might be some other data and model limitations. We droped certain amount of data who don't have LDL-C value records in our data. Patients might also purchased statins in a way that cannot be tracked by the pharmacy claims. Multiple causal related model assumptions including no unmeasured confounding, positivity, generalizability, and clinical equipoise also restricted the practical use of our PSTP.

5. Conclusions

In conclusion, this new pipeline model of counterfactual prediction for statin medications using EHR and Claims data provides a valuable approach to predict clinical outcomes for patients who are taking statins. It also thoroughly answered the major three research

questions proposed in the introduction. By leveraging the principles of weighting methods for counterfactual prediction, this model can incorporate both the patient's baseline information and their medical history in a robust and interpretable manner. Furthermore, it added the benefit aspect (LDL-C reduction) to the optimization modeling. Finally, we have CTS as the evaluation of the PSTP. The use of this model can help healthcare providers make more informed decisions about prescribing statin medications, leading to improved patient outcomes and more efficient use of resources. It is a promising area for further research and development in the field of machine learning for healthcare applications.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank the support from NIH 1R01HL143390-01A1 and help from OptumLabs staff for the data extraction.

References

[1]. Thompson PD, Parker B. Statins, exercise, and exercise training. J Am Coll Cardiol Aug. 2013;62(8):715–6. 10.1016/j.jacc.2013.03.030.

[2]. Thompson PD, Panza G, Zaleski A, Taylor B. Statin-associated side effects. J Am Coll Cardiol 2016;67(20):2395–410. [PubMed: 27199064]

[3]. Sun B, et al. Development and application of pharmacological statin-associated muscle symptoms phenotyping algorithms using structured and unstructured electronic health records data. medRxiv 2023;2023(05):9–10.

[4]. Chi C-L, et al. Producing personalized statin treatment plans to optimize clinical outcomes using big data and machine learning. J Biomed Inf Apr. 2022;128: 104029. 10.1016/j.jbi.2022.104029.

[5]. Li F, Li F. Propensity score weighting for causal inference with multiple treatments. Ann Appl Stat Dec. 2019;13(4). 10.1214/19-AOAS1282.

[6]. Penson PE, Pirro M, Banach M. LDL-C: lower is better for longer—even at low risk. BMC Med 2020;18:1–6. [PubMed: 31898501]

[7]. Caramia M, Dell'Olmo P. Multi-objective optimization. In: Caramia M, Dell'Olmo P, editors. Multi-objective management in freight logistics: increasing capacity, service level, sustainability, and safety with optimization algorithms. Cham: Springer International Publishing; 2020. p. 21–51. 10.1007/978-3-030-50812-8_2.

[8]. Malarvizhi R and Thanamani DAS, "K-nearest neighbor in missing data imputation"..

[9]. McCaffrey DF, Griffin BA, Almirall D, Slaughter ME, Ramchand R, Burgette LF. A tutorial on propensity score estimation for multiple treatments using generalized boosted models. Stat Med Aug. 2013;32(19):3388–414. 10.1002/sim.5753. [PubMed: 23508673]

[10]. Collier ZK, Leite WL, Karpyn A. Neural networks to estimate generalized propensity scores for continuous treatment doses. Eval Rev Mar. 2021:0193841X2199219. 10.1177/0193841X21992199.

[11]. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, and Salakhutdinov R, "Dropout: a simple way to prevent neural networks from overfitting"..

[12]. Li F, Morgan KL, Zaslavsky AM. Balancing covariates via propensity score weighting. J Am Stat Assoc Jan. 2018;113(521):390–400. 10.1080/01621459.2016.1260466.

[13]. Tripepi G, Chesnaye NC, Dekker FW, Zoccali C, Jager KJ. Intention to treat and per protocol analysis in clinical trials. Nephrology 2020;25(7):513–7. [PubMed: 32147926]

[14]. Tibshirani R Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B Methodol. 1996;58(1):267–88.

[15]. Wang Z, Rangaiah GP. Application and analysis of methods for selecting an optimal solution from the Pareto-optimal front obtained by multiobjective optimization. Ind Eng Chem Res 2017;56(2):560–74.

[16]. Austin PC. Using the standardized difference to compare the prevalence of a binary variable between two groups in observational research. Commun Stat Simulat Comput 2009;38(6):1228–34.

[17]. Stone NJ, et al. “2013 ACC/AHA guideline on the treatment of blood cholesterol to reduce atherosclerotic cardiovascular risk in adults: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines,”. Circulation 2014;129(25_suppl_2):S1–45. [PubMed: 24222016]

[18]. Nethery RC, Mealli F, Dominici F. Estimating population average causal effects in the presence of non-overlap: the effect of natural gas compressor station exposure on cancer mortality. Ann Appl Stat 2019;13(2):1242. [PubMed: 31346355]

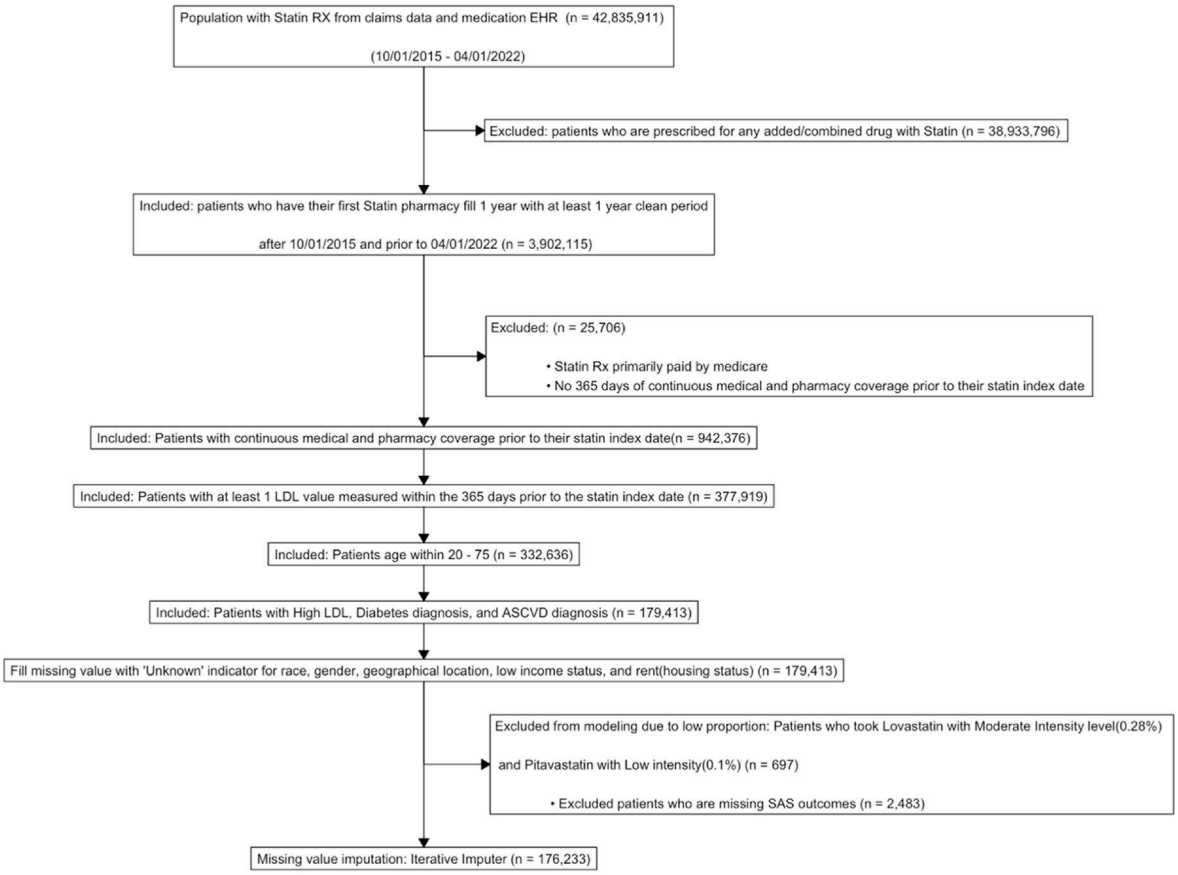


Fig. 1.
Data extraction and preprocessing consort diagram.

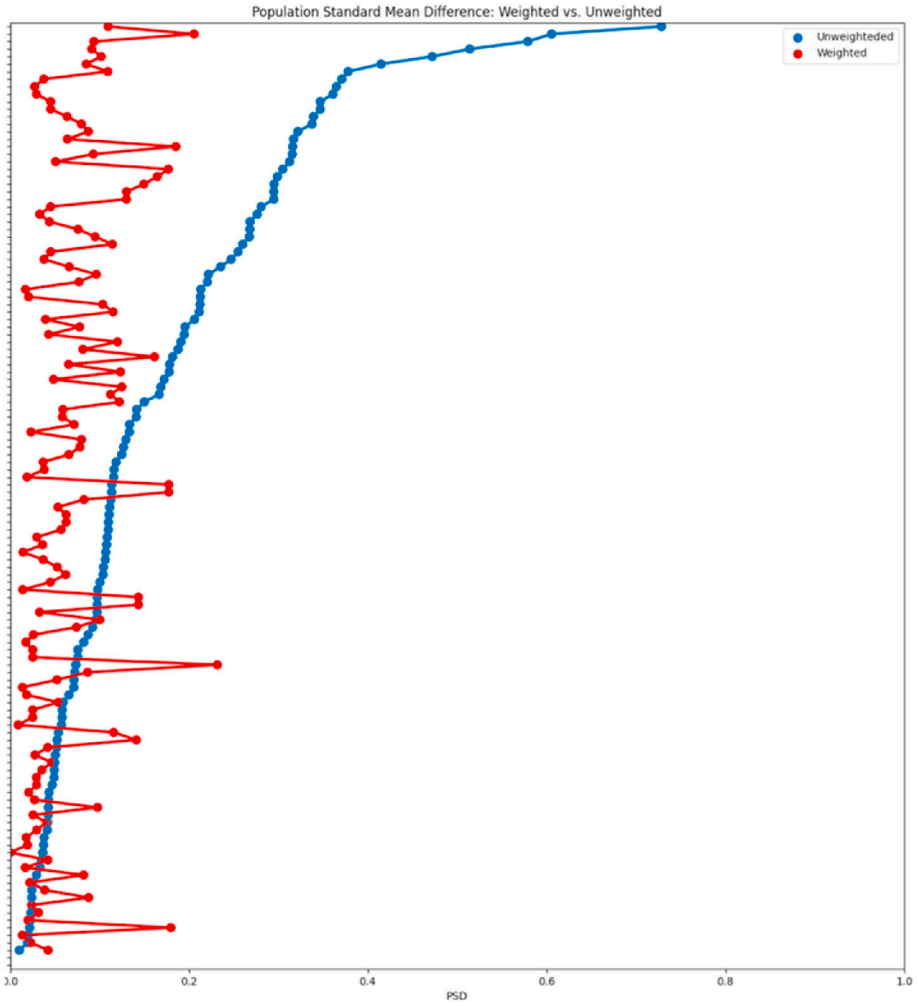


Fig. 2(a).
PSD of sorted unweighted covariates.

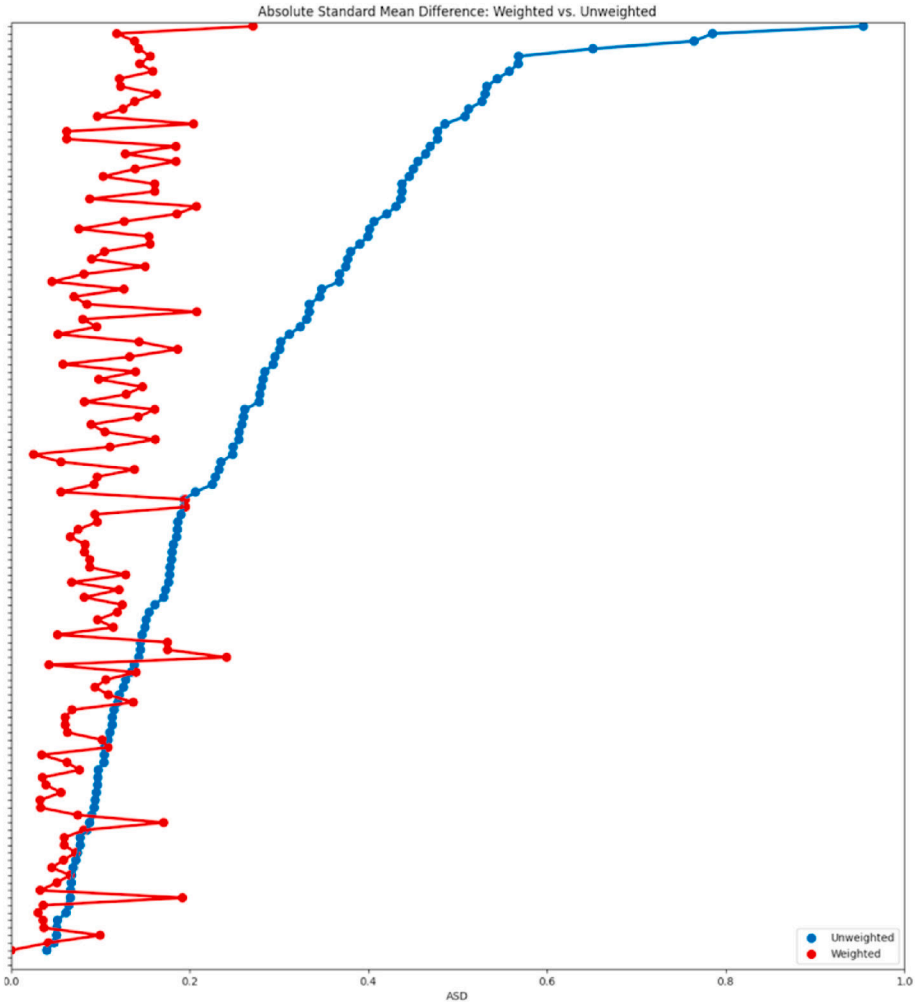


Fig. 2(b).
ASD of sorted unweighted covariates.

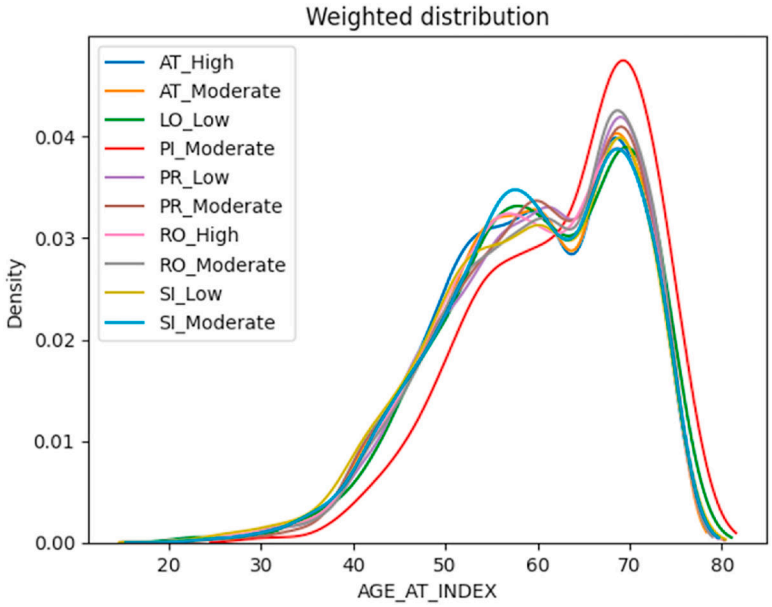


Fig. 3(a).
Weighted distribution of patients' age at index date.

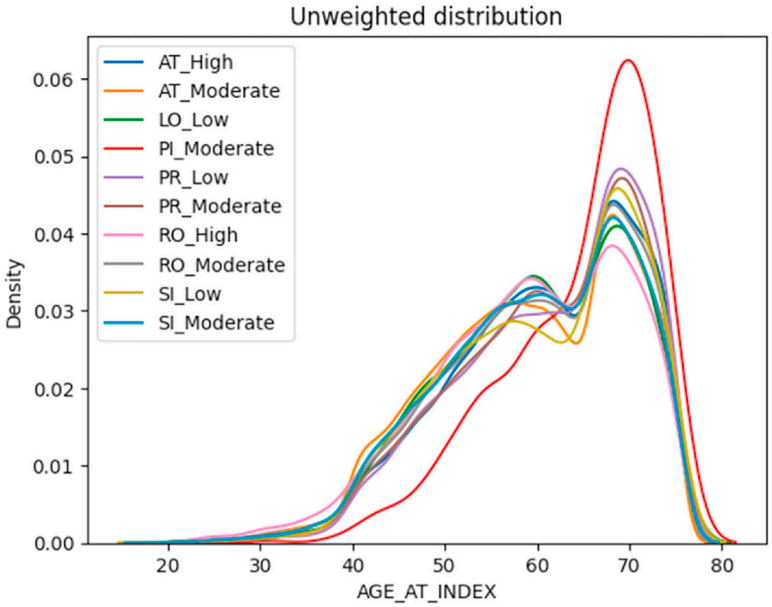


Fig. 3(b).
Unweighted distribution of patients' age at index date.

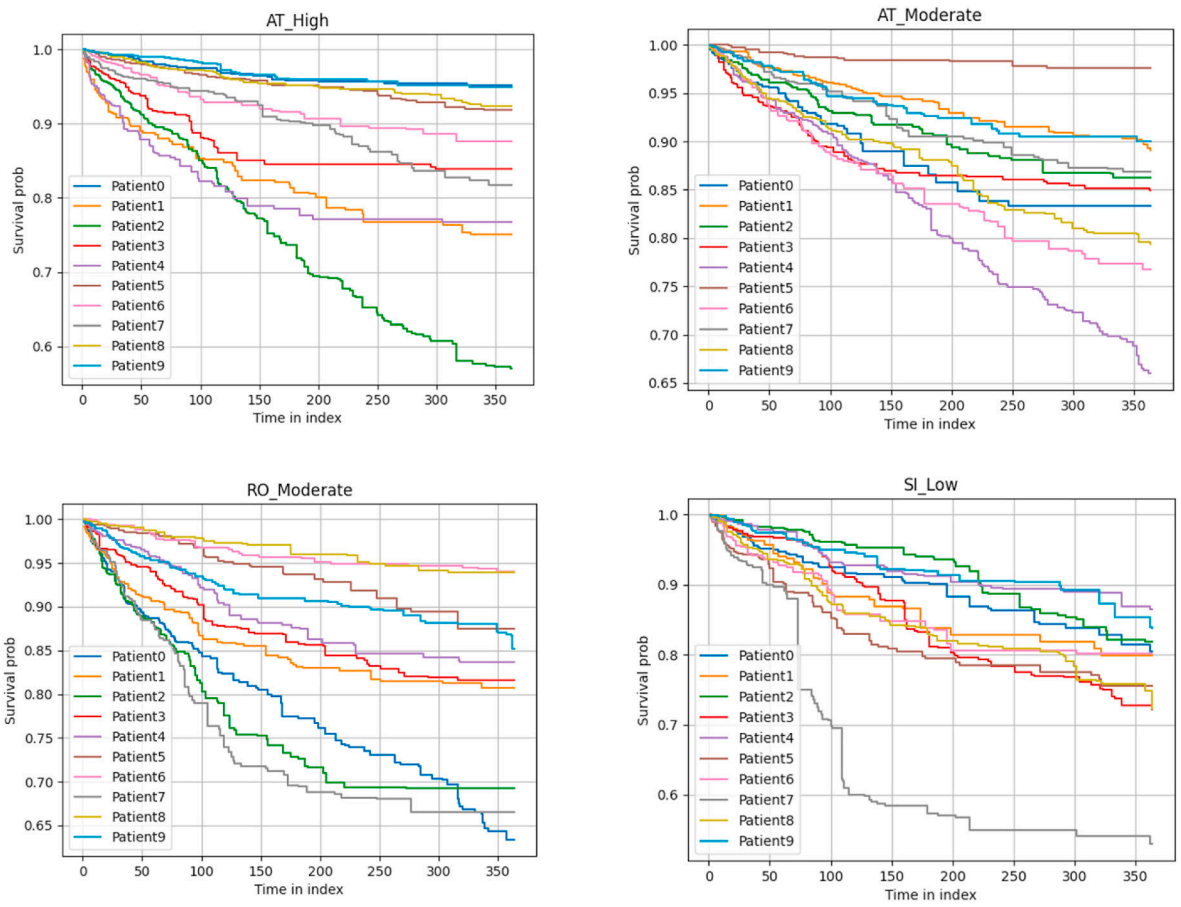


Fig. 4.

Counterfactual survival plot of SAS for randomly simulated patients for Atorvastatin High/ Moderate intensity and Simivastatin Low intensity and Rosuvastatin Moderate intensity.

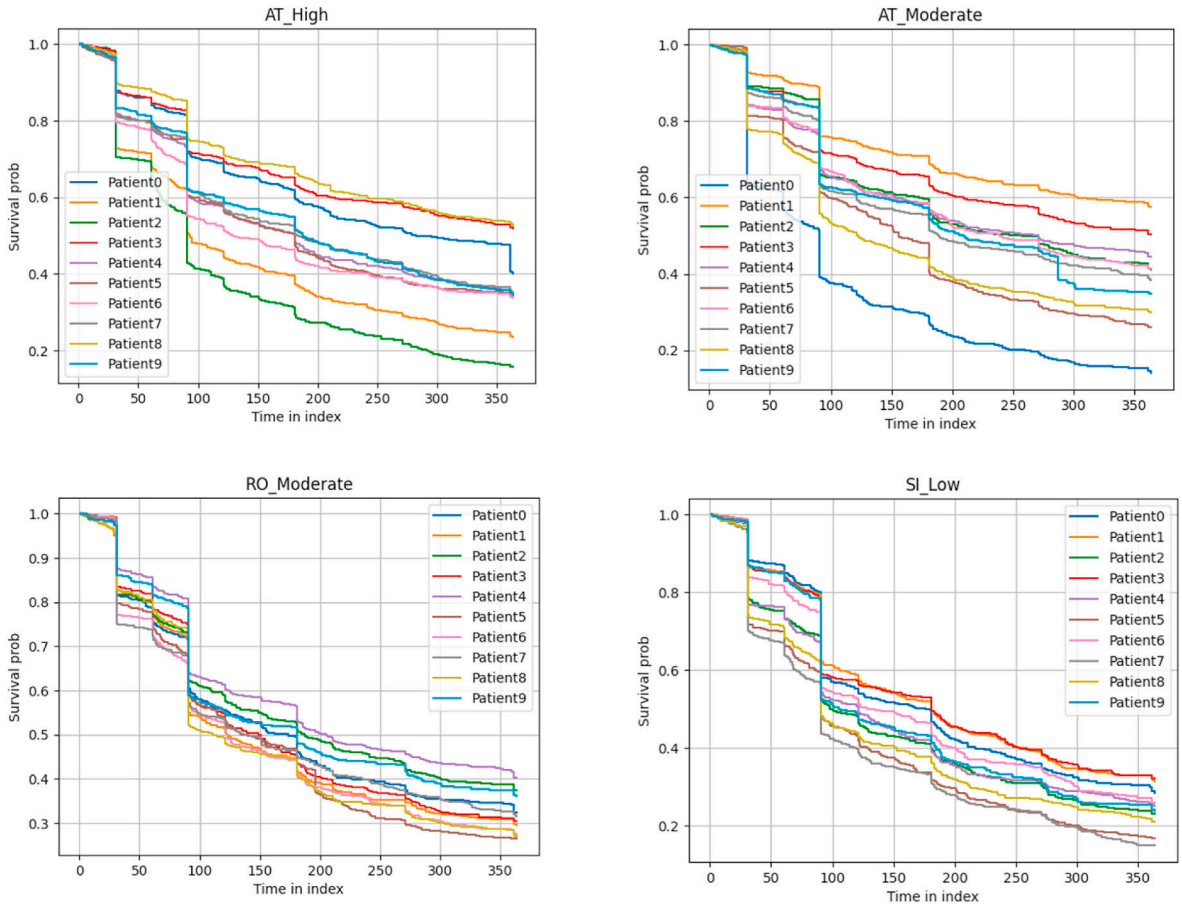
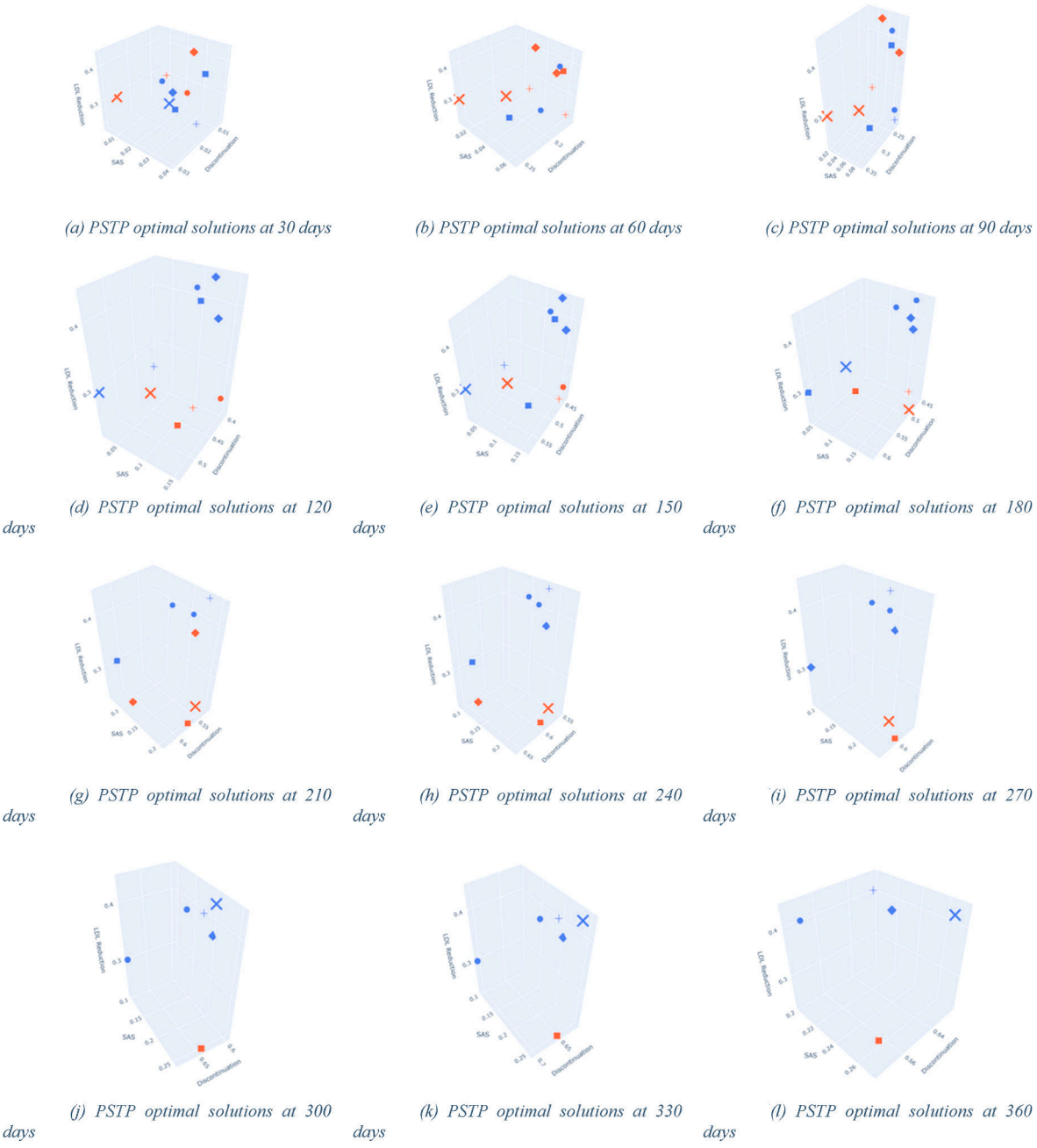


Fig. 5.

Counterfactual survival plot of Discontinuation for randomly simulated patients for Atorvastatin High/Moderate intensity and Simivastatin Low intensity and Rosuvastatin Moderate intensity.



(a) PSTP optimal solutions at 30 days (b) PSTP optimal solutions at 60 days (c) PSTP optimal solutions at 90 days

(d) PSTP optimal solutions at 120 days (e) PSTP optimal solutions at 150 days (f) PSTP optimal solutions at 180 days

(g) PSTP optimal solutions at 210 days (h) PSTP optimal solutions at 240 days (i) PSTP optimal solutions at 270 days

(j) PSTP optimal solutions at 300 days (k) PSTP optimal solutions at 330 days (l) PSTP optimal solutions at 360 days

Fig. 6.

Example plot of 3D Pareto Optimization for selecting out PSTP treatment for a simulated patient for 30 days–360 days survival risks of SAS, discontinuation, and LDL-C reduction. Purple dot means that it is pareto front options. Orange dot is worse choice.

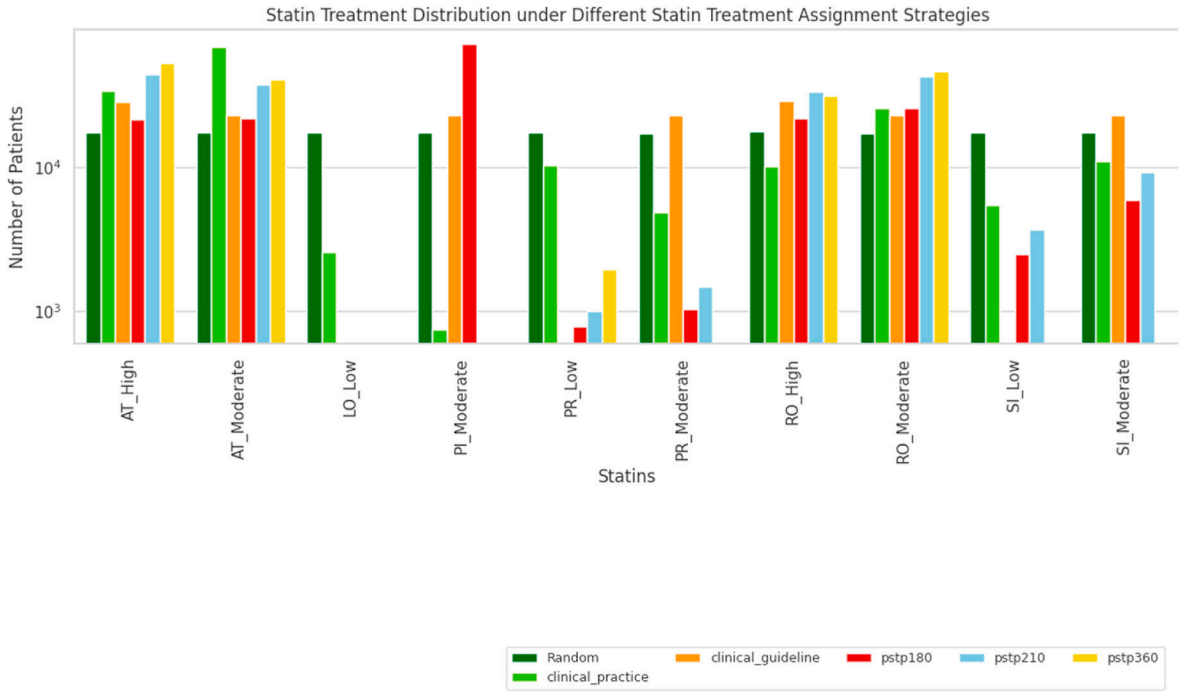


Fig. 7.
Summarization of the CTS result with multiple survival days for PSTP predictions result.

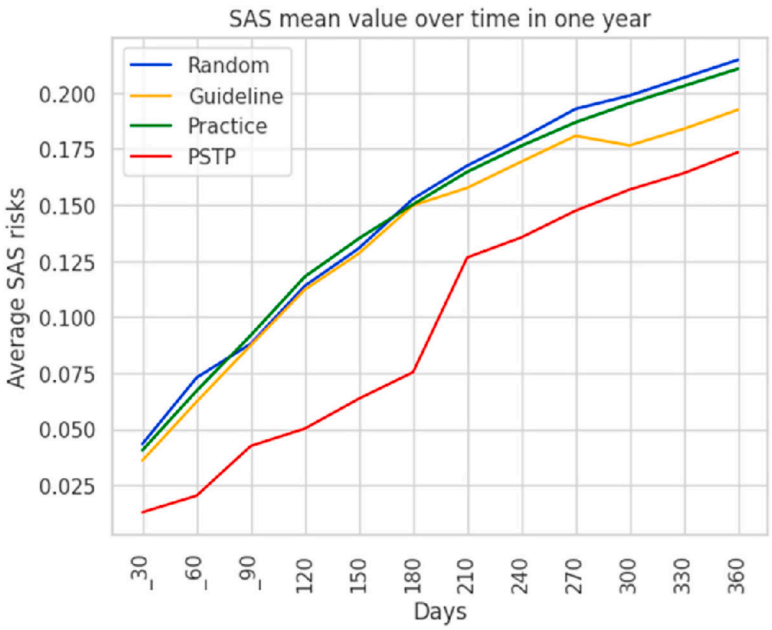


Fig. 8(a).
Summarized SAS survival risks over time for each CTS arm.

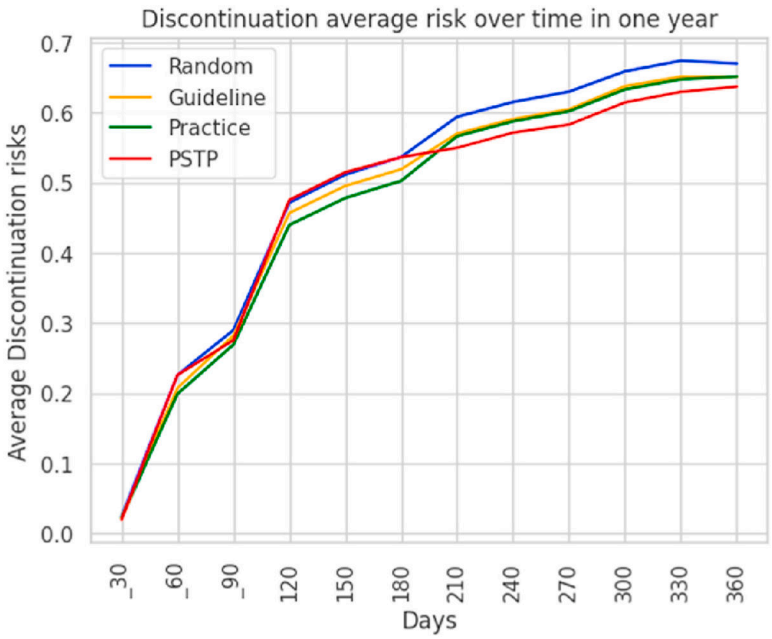


Fig. 8(b).
Summarized Disc survival risks over time for each CTS arm.

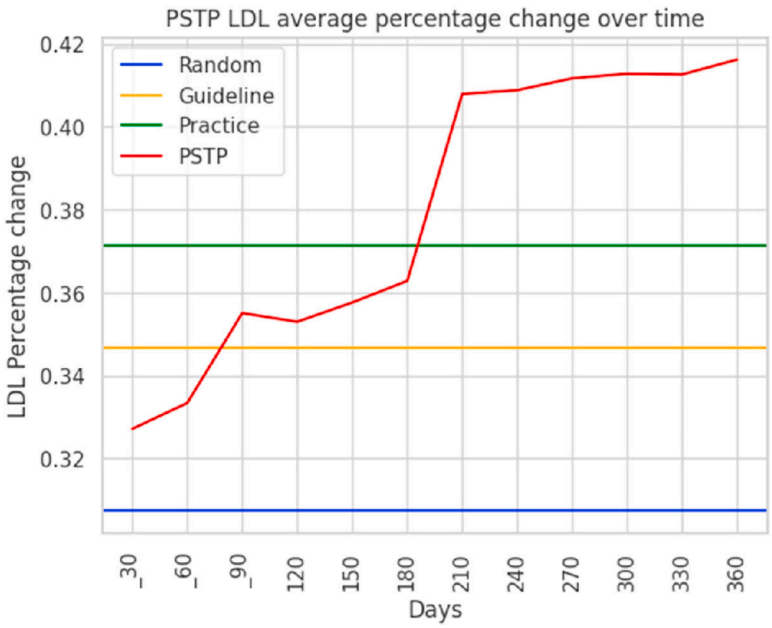


Fig. 8(c).
Summarized LDL-C percentage reduction value over time for each CTS arm.

Table 1

Statin dosage intensity groupings.

Statin generic\Intensity level	Low	Moderate	High
Atorvastatin	–	10 mg, 20 mg	40 mg, 80 mg
Lovastatin	10 mg, 20 mg	40 mg	60 mg
Pitavastatin	1 mg	2 mg, 4 mg	–
Pravastatin	10 mg, 20 mg	40 mg, 80 mg	
Rosuvastatin	–	5 mg, 10 mg	20 mg, 40 mg
Simvastatin	5 mg, 10 mg	20 mg, 40 mg	80 mg

Table 2

Subset of Baseline covariates across different Statin types(each cell is mean(SD) or proportion of each column; Total N is defined in first row).

	Atorvastatin	Lovastatin	Pitavastatin	Pravastatin	Rosuvastatin	Simvastatin	Overall
N	104551	2569	736	15240	36505	16632	176233
Intensity Level High	34567	0	0	0	10281	0	44848
Intensity Level Moderate	69984	0	736	4872	26224	11101	112917
Intensity Level Low	0	2569	0	10368	0	5531	18468
Baseline Characteristics:							
Age at index	59.8 (10.2)	60.3 (9.6)	64.9 (8.0)	61.3 (9.7)	59.9 (10.2)	60.1 (10.0)	60.0 (10.1)
Average LDL value prior index	132.7 (42.4)	128.8 (39.1)	147.4 (44.3)	130.4 (39.1)	144.1 (46.3)	130.2 (41.2)	134.6 (43.1)
Age adjusted Charlson score	4.0 (2.7)	3.9 (2.4)	4.8 (2.8)	4.2 (2.6)	3.7 (2.6)	3.9 (2.6)	4.0 (2.7)
Charlson Quan score	2.4 (2.3)	2.3 (2.0)	2.7 (2.5)	2.5 (2.3)	2.1 (2.2)	2.3 (2.2)	2.4 (2.3)
Charlson Quan updated score	1.3 (1.8)	1.1 (1.5)	1.5 (1.8)	1.3 (1.7)	1.1 (1.6)	1.2 (1.6)	1.3 (1.7)
Elixhauser ahrq score	1.4 (9.0)	−0.2 (7.5)	1.3 (8.1)	0.7 (8.3)	0.5 (7.7)	0.1 (7.8)	1.0 (8.6)
Elixhauser van score	3.7 (7.9)	2.2 (6.3)	3.8 (7.3)	3.1 (7.2)	2.8 (6.8)	2.5 (6.9)	3.3 (7.5)
Smoking never used	7.9%	7.3%	7.3%	8.0%	9.3%	7.5%	8.2%
Smoking previously used	25.8%	19.8%	20.6%	21.8%	20.7%	20.2%	23.8%
Smoking currently using	1.7%	1.3%	<1.2%	1.3%	1.4%	1.5%	1.6%
Medication at index:							
Rx at index lipid regulating	3.6%	4.0%	16.0%	5.3%	5.7%	4.0%	4.3%
Rx at index immunosuppressant	1.5%	1.6%	2.0%	2.2%	1.7%	1.3%	1.6%
Rx at index blood pressure med	63.9%	64.0%	68.8%	66.1%	57.1%	61.2%	62.5%
Medications prior index date:							
Rx prior year lipid regulating	5.7%	6.8%	27.8%	8.4%	8.7%	6.2%	6.7%
Rx prior year immunosuppressant	2.8%	2.9%	4.4%	3.7%	3.1%	2.4%	2.9%
Rx prior year blood pressure med	70.3%	71.5%	75.1%	73.5%	64.2%	68.3%	69.2%
Race:							
Asian	3.7%	3.0%	2.6%	2.9%	3.5%	4.4%	3.7%
Black	13.2%	14.2%	12.2%	15.8%	11.0%	13.9%	13.0%
Hispanic	13.7%	18.0%	9.0%	12.4%	12.5%	18.2%	13.8%
White	50.0%	43.5%	55.3%	49.0%	54.5%	43.3%	50.1%

	Atorvastatin	Lovastatin	Pitavastatin	Pravastatin	Rosuvastatin	Simvastatin	Overall
Gender:							
Male	47.5%	54.9%	63.5%	57.0%	50.6%	51.9%	49.6%
Gender unknown	1.3%	1.2%	<1.2%	0.9%	1.7%	1.3%	1.3%
Geographical location:							
Midwest	20.4%	16.7%	12.1%	14.9%	12.1%	14.9%	17.6%
Northeast	11.2%	4.7%	10.5%	7.6%	13.3%	11.4%	11.2%
Other	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
South	53.7%	65.2%	67.0%	67.5%	61.8%	60.2%	57.4%
West	13.2%	11.8%	9.5%	8.9%	10.9%	11.9%	12.2%
Income status:							
Level 1.0	7.6%	9.1%	9.8%	8.9%	5.6%	8.9%	7.4%
Level 2.0	3.3%	4.0%	3.2%	3.8%	2.0%	3.8%	3.2%
Level 3.0	0.9%	0.8%	<1.2%	0.8%	0.4%	0.9%	0.8%
Level 4.0	0.6%	0.9%	<1.2%	0.8%	0.5%	0.7%	0.6%
Level unknown	56.1%	53.8%	34.3%	48.4%	58.8%	53.8%	55.6%