Feature Selection in Predictive Modeling: A Systematic Study on Drug Response Heterogeneity for Type II Diabetic Patients

Jingyuan Chou, James Flory, Fei Wang*
Department of Healthcare Policy and Research. Weill Cornell Medicine.
***Corresponding Author. Email:few2001@med.cornell.edu**

Abstract

With the rapid development of computer hardware and software technologies, more and more electronic health data from insurance claims, clinical trials and hospitals are becoming readily available. These data provide a rich resource for developing various healthcare analytics algorithms, among which predictive modeling is of key importance in many real health problems. One important issue for data-driven predictive modeling is high dimensionality, and feature selection is one effective strategy to reduce the number of independent variables and control the confounding factors. However, most of the existing studies just pick one feature selection approach without comprehensive investigations. In this paper, we investigate the issue of drug response heterogeneity for type II diabetes mellitus (T2DM) patients using a large scale clinical trial data. Our goal is to find out the important factors that may lead to the response heterogeneity for three popular T2DM drugs, Metformin, Rosiglitazone and Glimepiride. We implemented 8 different feature selection approaches and compared their performances with various measures including prediction error and the consistency of the identified important factors. Finally, we ensemble all factor lists picked by different algorithms and obtain a final set of factors that contribute to the drug response heterogeneities and verified them through existing literature.

1 Introduction

Diabetes is a problem with the human body that causes high blood glucose, which is also known as hyperglycemia. Type 2 Diabetes Mellitus (T2DM) is the most common form of diabetes when the human body has insulin resistance. According to statistics, Diabetes affects more than 285 million people globally[1] with 90% of the cases diagnosed as T2DM[2].

T2DM is a progressive disease with a consistent and steady increase in glycosylated hemoglobin (HbA1c) over time. T2DM patients are also associated with enhanced risk of micro- and macrovascular complications and a substantial reduction in life expectancy. Three major pathophysiologic abnormalities associated with T2DM are impaired insulin secretion, excessive hepatic glucose output, and insulin resistance in skeletal muscle, liver, and adipose tissue[3]. These defects have been treated in clinical praxis by use of oral insulin secretagogues (sulfonylureas (SU)/glinides) or insulin sensitizers (metformin and thiazolidinediones (TZDs)) respectively. Rosiglitazone is a insulin sensitizer in TZD family and glimepiride is an insulin secretagogue in the SU family. Rojas and Gomes mentioned[4] that metformin/glimepiride combination results in a lower HbA1c concentration and fewer hypoglycemic events when compared to the glibenclamide/metformin combination. Metfomin/SU combination therapy was also associated with reduced all-cause mortality. Studies have respoted that the use of rosiglitazone was associated with a 5%, non-significant, reduction in mortality[5], and Raef *et al*. found that[6] rosiglitazone, when added to Metformin in type 2 DM patients, was effective and well tolerated.

T2DM is a heterogeneous disease with large variation in the relative contributions of insulin resistance and beta cell dysfunction between subgroups and individuals[7]. The response to treatment for T2DM typically varies among individuals within a study population, which is known as heterogeneity of treatment response[8]. Many studies have been reported for the investigation of the treatment response heterogeneity of T2DM patients, where the response to pharmacological treatment was typically measured by HbA1c and/or fasting plasma glucose (FPG) levels during follow-up. Many of these approaches are hypothesis driven, and one can refer to[8] for a comprehensive survey.

The goal of this study is to investigate the application of machine learning approaches in identifying important factors that contribute to the response heterogeneity of T2DM drugs. Different from conventional statistical hypothesis driven approaches, those machine learning algorithms are data-driven and hypothesis free. Three important drugs mentioned above, namely metformin, rosiglitazone and glimepiride, are investigated. The patient data are from the ACCORD trial[9].

From the language of machine learning, such factor identification problem can be tackled by feature selection approaches. Feature selection[10] refers to the process of selecting a small subset of features from the entire feature set according to some criteria. Typically feature selection approaches can be categorized as either filter methods or wrapper methods. The filter methods typically calculate a score (e.g., correlation with target variables) for each feature and then rank the features according to the scores and pick the leading ones. Wrapper methods integrate the feature selection process with a learner e.g., a classifier such as Support Vector Machines and the selected features are guaranteed to lead to better performance of the learner. Different feature selection algorithms have different assumptions and advantages. For example, filter methods are computationally efficient, while wrapper methods are coupled with endpoint learner and thus can lead to better performance. Therefore picking an appropriate feature selection algorithm is not a trivial task for different applications.

In our study, we systematically investigate the different feature selection algorithms in the context of response heterogeneity of T2DM drugs with ACCORD trial data. We compared the prediction performance using the features identified from those approaches to the response variable (change in HbA1c value), as well as the consistency of the feature sets selected by different algorithms. Finally we come up with a strategy on ensemble the feature sets picked by different approach and verify them with evidence from existing literatures.

2 Methods

The detailed methodologies we used in our research are introduced in this section.

2.1 Data

The data we used in our investigation are from Action to Control Cardiovascular Risk in Diabetes (ACCORD) database, time-series data from baseline to follow up to 7 years, 84 months, for 10251 patients, each patient is under intensive glycemia control or standard glycemia control. All the patients are grouped into 8 arms, each patient only belongs to one arm, the time unit of record is month, in our analysis, subset of 10251 patients were used as input.

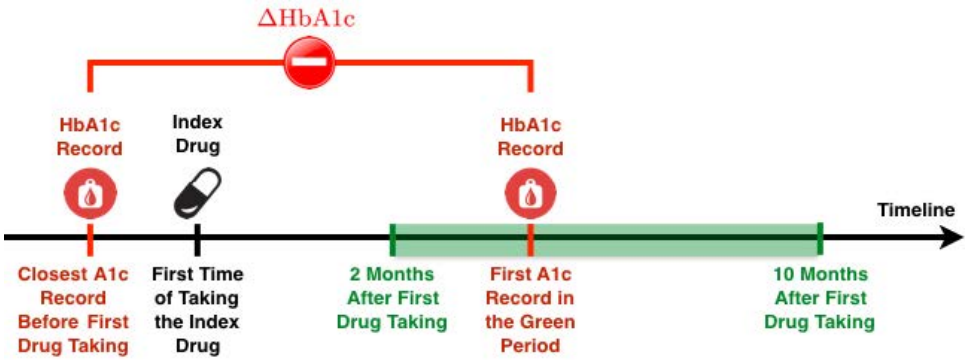

Figure 1: The graphical illustration of our setup on how to calculate ΔHbA1c

Three cohorts were initially constructed in our study, which are the patients who took metformin, rosiglitazone and glimepiride (we call these drugs index drugs). To reduce the effects of combination therapies, we require the first time the patient take the index drug should be before the patient takes insulin or any other drugs functioning similarly to the index drug. We also exclude patients who took any T2DM drugs in three months before the first time they took the index drugs. This results in 521 patients in the metformin cohort, 1,127 patients in the rosiglitazone cohort and 688 patients in glimepiride cohort. The difference between the HbA1c values on baseline and follow up time points for each patient is used as the target. The baseline timestamp is set as the HbA1c value closest before the index drug taking time. The follow up timestamp is set as the earliest HbA1c value appeared between 2 months and 10 months after taking the index drugs. Figure 1 provides a graphical illustration on the details of our setup.

To compare the response heterogeneity measured by HbA1c change through the duration under different glycemic control (i.e., intensive or standard), we chose to split each drug cohort into 2 subgroup, rendering 6 cohorts in total for

further feature selection algorithms employment: metformin & intensive, metformin & standard, glimepiride & intensive, glimepiride & standard, rosiglitazone & intensive, rosiglitazone & standard. In the next following experiments, we would apply multiple feature selection algorithms on the top of the six cohorts to compare the performance and the consistency of them. The overview information about datasets are summarized in Table 1 and Table 2.

We selected patients who take the drug and set their baseline HbA1c as the latest record right or before the date they take the medicine, and set the first HbA1c record in the duration that from 2 month after they take the drug to 10 months after they take the drug as the follow-up HbA1c record. If someone who take several drugs together, they will belong to the cohort that the first drug they take, excluding those who take the second drug in 3 months after they take the first drug, thus avoiding the inter-related effect of combination of drugs

Table 1: Cohorts used in the study

Feature	Metformin		Glimepiride		Rosiglitazone	
	Intensive	Standard	Intensive	Standard	Intensive	Standard
#Sample	201	320	366	322	557	570
#Feature	139	135	140	138	140	138

Table 2: Mean and Std For Important Features In Multiple Datasets At Baseline

Feature	Metformin		Glimepiride		Rosiglitazone	
	Intensive	Standard	Intensive	Standard	Intensive	Standard
LDL	115.97 (39.05)	107.84 (34.71)	104.54 (34.84)	101.08 (34.27)	101.01 (31.13)	92.91 (30.52)
VLDL	39.77 (36.32)	37.48 (24.3)	36.01 (22)	41.44 (29.91)	33.47 (19.12)	40.02 (44.76)
HDL	41.19 (11.75)	41.77 (10.78)	42.8 (11.05)	41.69 (10.46)	42.74 (12.14)	40.92 (11.11)
Trig	212.5 (257)	199.8 (162.37)	185.81 (124.68)	224.17 (200.63)	173.71 (114.59)	211.54 (262.25)
BMI	31.88 (5.66)	31.52 (5.38)	31.57 (5.64)	32.14 (5.23)	30.93 (5.09)	31.54 (5.15)
Female	42%	33%	39%	41%	38%	34%
Age	62.78 (6.84)	62.96 (6.68)	62.24 (6.58)	61.13 (6)	62.69 (6.86)	61.95 (6.21)
Chol	196.91 (49.39)	187.08 (41.36)	183.35 (43.24)	184.21 (46.32)	177.22 (38.06)	173.86 (53.15)
SBP	135.27 (17.39)	132.33 (16.6)	134.04 (16.79)	132.29 (18.08)	132.52 (16.77)	130.9 (17)
DBP	76.85 (10.59)	74.65 (10.69)	75.88 (10.78)	75.41 (10.23)	73.97 (10.56)	73.39 (10.13)
HR	70.85 (11.79)	70.21 (11.03)	72.05 (11.05)	73.51 (11.8)	71.59 (11.79)	72.39 (11.33)
Yrsdiab	5.6 (6.06)	6.07 (5.06)	5.8 (5.88)	6.47 (5.36)	7.12 (6.06)	8.76 (6.31)
Waist cir	104.71 (14.53)	105.06 (13.56)	104.37 (14.84)	106.03 (12.96)	103.59 (13.27)	105.57 (13.17)
Weight	92.54 (19.09)	92.22 (18.48)	90.95 (18.97)	92.81 (17.4)	89.45 (17.06)	92.02 (18.39)
Height	170.18 (9.99)	170.86 (10.23)	169.49 (9.24)	169.82 (9.5)	169.91 (9.47)	170.49 (9.53)
FPG	167.12(58.94)	181.09(51.75)	160.64(48.49)	189.96(58.68)	149.81(47.4)	188.11(51.04)
GFR	92.82 (23.51)	89.1 (23.76)	91.02 (24.28)	89.2 (25.86)	89.29 (26.74)	88.04 (22.53)
SCREAT	0.87 (0.21)	0.91 (0.21)	0.89 (0.23)	0.91 (0.28)	0.92 (0.24)	0.93 (0.25)
HbA1c	7.85 (1.15)	8.26 (1.05)	7.75 (1.2)	8.42 (1.02)	7.38 (1.16)	8.46 (0.95)

Number in the parenthesis is the standard deviation; LDL: Low density lipoprotein (mg/dL); VLDL: Very low density lipoprotein (mg/dL); HDL: High density lipoprotein (mg/dL); Trig: Triglycerides (mg/dL); BMI: Body Mass Index; Female: Gender; Age: Age at baseline; Chol: Total Cholesterol (mg/dL); SBP: Systolic Blood Pressure (mmHg); DBP: Diastolic Blood Pressure (mmHg); HR: Heart Rate (bpm); Yrsdiab: Years of diabetes; Waist cir: waist circumference (cm); Weight: Kg; Height: cm; FPG: Fasting plasma glucose (mg/dL); GFR: eGFR from 4 variable MDRD equation (ml/min/1.73 m2); SCREAT: Serum creatinine (mg/dL); HbA1c: Hemoglobin A1C

2.2 Experimental Setup

The goal of our experiment is to compare the feature selection methods on different cohorts, and also compare the difference of the result of feature selection on both intensive and standard cohort of specific drug.

The outcome of our study is the HbA1c change through the duration, which is a continuous value, indicating our model

employed later will be a regression model. The features for each cohort in table 1 vary, because we drop the features with 20% or more missings. Eventually, there are 135 to 140 features for each cohort, pretty much the same. Features include health status, medication log, demographic information, daily life, etc. There are continuous, binary and categorical features in the features of cohort, categorical features are converted to binary ones using one-hot encoding and we don't do any discretization to continuous features.

However, there are still features with less than 20% missing, we adopted multivariate imputation by chained equations (MICE)[11] to impute the missing feature values. Then the pre-processed data were fed into different feature selection algorithms. All of the work is done using R v3.4.3.

2.3 Feature Selection Algorithms

Table 3: Comparison between different feature selection approaches

Methods	Single/Multiple Feature	Filter/Wrapper	Complexity	Subset Generation	Stability	Used to Eliminate
Correlation Coefficient	Single	Filter	O(1)	Forward Selection	Stable	Irrelevant Features
Univariate LM	Single	Filter	O(N)	Weighted	Stable	Irrelevant features
MI	Single	Filter	O(NlogN)	Forward Selection	Stable	Redundant or Irrelevant Features
MRMR	Single	Filter	O(N^2)	Forward Selection	Stable	Redundant or Irrelevant Features
Multivariate LM	Multiple	Wrapper	O(C^2N)	Weighted	Not Stable	Irrelevant features
LASSO	Multiple	Wrapper	O(C^3+ C^2N)	Weighted	Not Stable	Irrelevant features
GBM	Multiple	Wrapper	Depends on tree	Weighted	Stable	Irrelevant features
Random Forest	Multiple	Wrapper	Depends on tree	Weighted	Stable	Irrelevant features

C denotes the number of features;

As we stated in the introduction, feature selection algorithms are typically categorized as filter or wrapper methods[10]. In the study, we investigated eight different feature selection methods, including both filter and wrapper methods: filter methods include correlation coefficient/statistical testing, univariate linear regression, Mutual Information (MI), Maximal Relevance Minimal Redundancy (mRMR)[12]. Wrapper methods include Multivariate linear regression, Least Absolute Shrinkage and Selection Operator (LASSO)[13], Gradient Boosting Machine (GBM)[14], Random Forest (RF)[15]. Detailed implementation of each method are described as follows.

Table 3 provides a comprehensive comparison of the characteristics of those different methodologies in terms of theoretical algorithm. In our work, we will evaluate the consistency and the predictive ability for all the algorithms.

Correlation based approach evaluate the importance of each feature with response variable by calculating their correlations testing. In our case, the response variable is the change in HbA1c value. Therefore if the feature is continuous, Pearson/Spearman Correlation test is performed, if the feature is categorical or binary, ANOVA/Kruskal-Wallis test/Wilcoxon signed-rank test is applied. The p-value and correlation coefficient from the statistical model are used to rank each feature. Note that before we do the actual test, we first check if the response value distribution and the feature distribution are both normal, if at least one of them is not normal, nonparametric tests are employed to check the dependency, otherwise parametric tests are performed.

Univariate linear regression(ULR)[16] is a typical generalized linear model, we considered a feature as input at a time, and HbA1c change as the label for the univariate model. Thus, there will be a linear regression model for each feature, which outputs a coefficient and also a p value for the corresponding feature, the p value will be the proof for its ranking among all features in the method. The smaller the p value, the higher the rank.

Mutual information(MI)[17] the mutual dependence between the two random variables, it measures the information that two features share: how much knowing one of these variables reduces uncertainty about the other.MI is used as a feature selection approach in Natural Language Processing area, especially search engine. The R package $infotheo$ and $mpmi$ provides functions to calculate mutual information efficiently, each variable has mutual information with HbA1c change.

Maximal Relevance Minimal Redundancy(mRMR)[12] is also a feature selection algorithm frequently used in various application. mRMR is shown to be more powerful than simple forward or backward selection, because it also evaluates features that are mutually independent but still have high correlation or dependency with the response variable. In mRMR, the feature redundancy and feature/response dependency are measured by mutual information. We implemented this using the package $mRMRe$ in R.

Multivariate Linear Regression (MLR)[18] is the most common model for ordinary least square regression problem, which aims to minimize the least square error between the linear prediction value and the true label. MLR can also be utilized to select a subset of important features which are correlated to the label, whatever the direction that they have influence on the label. Similarly, the multivariate linear regression model can also output a coefficient and p value for each feature. The coefficient can be viewed as the weight for that feature, and p value indicates if the weight, which suggests the contribution of the corresponding feature to the final prediction of the response, is significant. In our implementation, all the features except HbA1c change are treated as independent variables, and HbA1c change is the response variable. The model is implemented in 10 fold cross validation.

Least Absolute Shrinkage and Selection Operator (LASSO)[13] is an extension of multivariate linear regression with a ℓ_1 norm regularizer penalizing the sparsity of the coefficient vector. In our implementation, we set change of HbA1c as target and all other features except HbA1c change as input, $glmnet$ package in R was used, 10 fold cross-validation were employed, hyperparameter λ was tuned from 0 to 10 , 0.001 as stepsize.

Gradient Boosting Machine (GBM)[14] is one of the most popular and effective ensemble supervised learning methods, it constructs a forward stage-wise additive model by implementing gradient descent in functional space. GBM computes the feature importance based on the number of times a variable is selected for splitting, weighted by the squared improvement to the model as a result of each split, and averaged over all trees.[19] In our implementation, we used $caret$ package in R, we set HbA1c change as the label, and other features as the input, we evaluated the following model hyperparameters are tuned with 10 fold cross-validation: the number of trees, from 100 to 1000, 100 trees as step size; shrinkage parameter,from 0.001 to 0.6, 0.001 as step size; Minimum observation in each node, 10, and the depth of the tree, from 4 to 12, 1 as step size. The optimal set of parameters varies depending on the data.

Random Forest (RF)[15] is an ensemble of decision trees with the bagging strategy. It usually achieves accurate and stable prediction results. RF measures the features importance by impurity index, which is the total decrease in node impurities measured by Gini Index from splitting with the variable, averaged over all trees. The $randomForest$ package in R provides function for implementation. In the experiment, the number of trees is tuned through 10-fold cross validation, the number of trees is mainly tuned from 10 to 1000, 10 as step size. The label is HbA1c change while other features as input, Root Mean Square Error(RMSE) is also the evaluation metric in our experiment.

As different feature selection methods derive different results, an effective measure to investigate the heterogeneity of those results is needed, which can indicate the most common and reliable feature selection methods. The most reliable feature selection method can reduce the heterogeneity of the selected features, which are more reasonable for further usage. In addition to prediction performance,we check the previous work[20] to find effective measures for comparison between multiple feature selection methods, and we derive the following consistency index to quantify the algorithm consistency.
Let $F = \{f_1,...,f_c\}$ be a set of c features, K be the number of features selected, and $M = \{M_1,...,M_k\}$ is a set of output feature lists of k algorithms, and M_{id},M_{jd} are the subset of features M_{id}, M_{jd} are subset of F,where the cardinality

Table 4: The Root Mean Square Errot(RMSE) of Wrapper Methods

Cohort	Multi lm	Lasso	GBM	RF
Metformin Int	1.119	**0.678**	0.697	0.697
Metformin Sta	1.253	0.894	**0.889**	0.926
Glimepiride Int	0.814	0.617	**0.596**	0.624
Glimepiride Sta	1.089	0.855	**0.852**	0.876
Rosiglitazone Int	0.685	0.629	**0.603**	0.630
Rosiglitazone Sta	1.117	0.956	**0.952**	0.958
Mean	1.013	0.772	**0.764**	0.785

of M_{id} and M_{jd} are d,then the consistency index can be calculated as

$$consistencyindex = \frac{2}{M}\sum_{i!=j}^{M}\frac{|M_{id} \cap M_{jd}| \cdot c - d^2}{d(c-d)} \tag{1}$$

3 Results

In this part, we evaluate utility and consistency of the features selected with different algorithms. The feature utilities are measured by their prediction performance, and we also use the proposed consistency index to measure the consistency of different feature selection approaches.

As the response variable, change in HbA1c value is continuous, we use Root Mean Square Error (RMSE) as the measure for prediction performance. Our dataset was first split into training (80%) and validation (20%) sets, and the algorithm hyperparameters were tuned via 10-fold cross validation on the training set. The results are summarized in Table 4 , where we only showed the performance of the 4 wrapper algorithms as the performance of filter methods will be dependent on the choice of predictors. From the table 4, we can observe that GBM always performs better than any other wrapper methods employed, except in Metformin intensive cohort, and Multivariate linear regression performs the worst. This is not surprising because of the nonlinear nature of GBM as well as its optimization based set up.

Table 5 lists the concrete consistency index values resulting from each algorithm in the six cohort with K = 10. From Table 5 we can observe that GBM, correlation testing, univariate linear regression and random forest are the methods are more stable than the other 4 methods. The consistency index for those 4 methods are pretty close, with a big gap compared with the other 4 unstable models.

Table 5: Consistency Index For Each Algorithm In Multiple Cohorts When Top 10 Features Are Selected

Cohort	Cor test	Uni lm	MI	MRMR	Multi lm	Lasso	GBM	RF
Metformin Int	0.45625	0.42929	0.34843	0.267578	0.4023437	0.29453	0.40234	**0.51015**
Metformin Sta	0.345161	**0.426209**	0.26411	0.102016	0.237096	0.399193	0.29112	0.318145
Glimepiride Int	**0.75329**	**0.75329**	0.45697	0.67248	0.160658	0.430038	0.69941	0.64554
Glimepiride Sta	**0.64429**	0.59035	0.50944	0.45551	0.401574	0.37460	**0.64429**	0.50944
Rosiglitazone Int	0.51085	0.45697	0.56472	0.26841	0.5916666	0.24147	**0.61860**	0.56472
Rosiglitazone Sta	0.509448	0.482480	0.42854	0.29370	0.2397637	0.320669	**0.59035**	0.45551
Mean	0.536	0.523	0.428	0.343	0.339	0.343	**0.5406**	0.50

Table 6 corresponds to the results with $K = 20$, which shows that the consistencies of all 8 methods decrease, however, GBM and statistical/correlation test are still fairly consistent compared to the others.

Figure 2 describes the consistency index for each algorithm in each cohort as the number of the features selected goes up, the number of features selected K varies from 10 to 30. The more stable the algorithm, the higher the consistency index. As the figure shows, Multivariate linear regression consistency is decreasing pretty fast on all cohorts. GBM always achieves high consistency values; RF is also relatively consistent. Correlation testing normally

Table 6: Consistency Index For Each Algorithm In Multiple Cohorts When Top 20 Features Are Selected

Cohort	Cor test	Uni lm	MI	MRMR	Multi lm	Lasso	GBM	RF
Metformin Int	0.44894	**0.55127**	0.30275	0.141949	0.1127118	0.25889	0.46355	0.34661
Metformin Sta	0.22192	0.25131	0.13377	0.1043859	0.01622	**0.31008**	**0.31008**	0.22192
Glimepiride Int	**0.68413**	0.47972	0.47972	0.47972	0.11890	0.34831	0.53813	0.45052
Glimepiride Sta	**0.41805**	0.38878	0.37414	0.25705	0.198504	0.31559	**0.41805**	**0.41805**
Rosiglitazone Int	0.61113	0.45052	0.55273	0.31911	0.18771	0.15850	**0.64033**	0.45052
Rosiglitazone Sta	0.13995	0.21314	0.28632	0.30096	0.02104	0.169230	**0.37414**	0.35950
Mean	0.421	0.389	0.354	0.267	0.105	0.26	**0.457**	0.372

Cor test: Correlation and Statistical test including Pearson correlation test, Spearman correlation test, ANOVA, Kruskal-Wallis Test, Mann-Whitney U test; Uni lm: univariate linear regression; MI: Mutual Information; MRMR: Maximal Relevance Minimal Redundancy; GBM: Gradient Boosting Machine; RF: Random Forest; Int: Intensive; Sta: Standard

shows high consistency when K is small and then decreases, except in Rosiglitazone intensive cohort. Univariate linear regression shows drastic change in all intensive cohorts, relatively smooth in standard cohorts.

It can be also noted that in datasets with most samples (Rosiglitazone standard), the range of consistency of all the feature selection methods become smaller. Most of the methods share the same tendency as the number of features go up. In general, it can be concluded that multivariate linear regression has the least consistency, GBM has the highest and the consistency of RF, correlation test and univariate linear regression have relatively consistency tendency, which is consistent with table 5 and table 6.

Therefore, in our setting, GBM, RF, Correlation testing, univariate linear regression are the consistent feature selection methods, which means that they tend to pick features consistent with other approaches. GBM and RF are wrapper methods and typically also achieve good prediction performance. Univariate linear regression and correlation testing are all filter methods. They are computationally more efficient but less consistent.

4 Discussion

In this part, we evaluated the qualitative analysis for our findings, which are consistent with the previous literature.

As Huang *et al.* suggest[21], they selected top 15 features as their best predictive features set, to make our results more informative, we select Top 20 features from output of consistent models (GBM, RF, Statistical Testing, Univariate linear regression) as our best predictive feature sets for each dataset. Each method has its own ranking, if there is a feature appears top 20 rank in at least two feature selection algorithm, the feature is selected to be one of the predictive features. Therefore, we have the following selected best feature sets for each cohort:

- For Metformin intensive cohort: hypoglycemic episodes, experience of shortness of breath, race(excluding Hispanic), had retinopathy, BMI, Cholesterol, HDL, Heart rate, height, weight, HbA1c level at baseline, FPG, GFR, Urinary albumin to creatinine ratio, Urinary creatinine, Urinary albumin, SBP, triglyceride, waist circumference. These 19 features are considered to be predictive according to the final ranking.
- For Metformin standard cohort: foot ulceration, number of oral agents taking, HbA1c at baseline, FPG, SBP, visual acuity score of left eye, VLDL, weight, waist circumference.
- For Glimepiride intensive cohort: Age, BMI, feeling score, gender, number of hypoglycemic episodes, HbA1c at base, HDL, height, GFR, triglyceride, waist circumference, weight.
- For Glimepiride standard cohort: Age, BMI, feeling score, vibration (perception at great toe), HbA1c at base, height, ALT, FPG, GFR, race(excluding hispanic), vitamin.
- For Rosiglitazone intensive cohort: Age, BMI, DBP, average frequency of blood sugar check, hypoglycemic episodes, HbA1c at base, LDL, FPG, GFR, Serum creatinine, race (black & white), SBP, triglyceride, vision

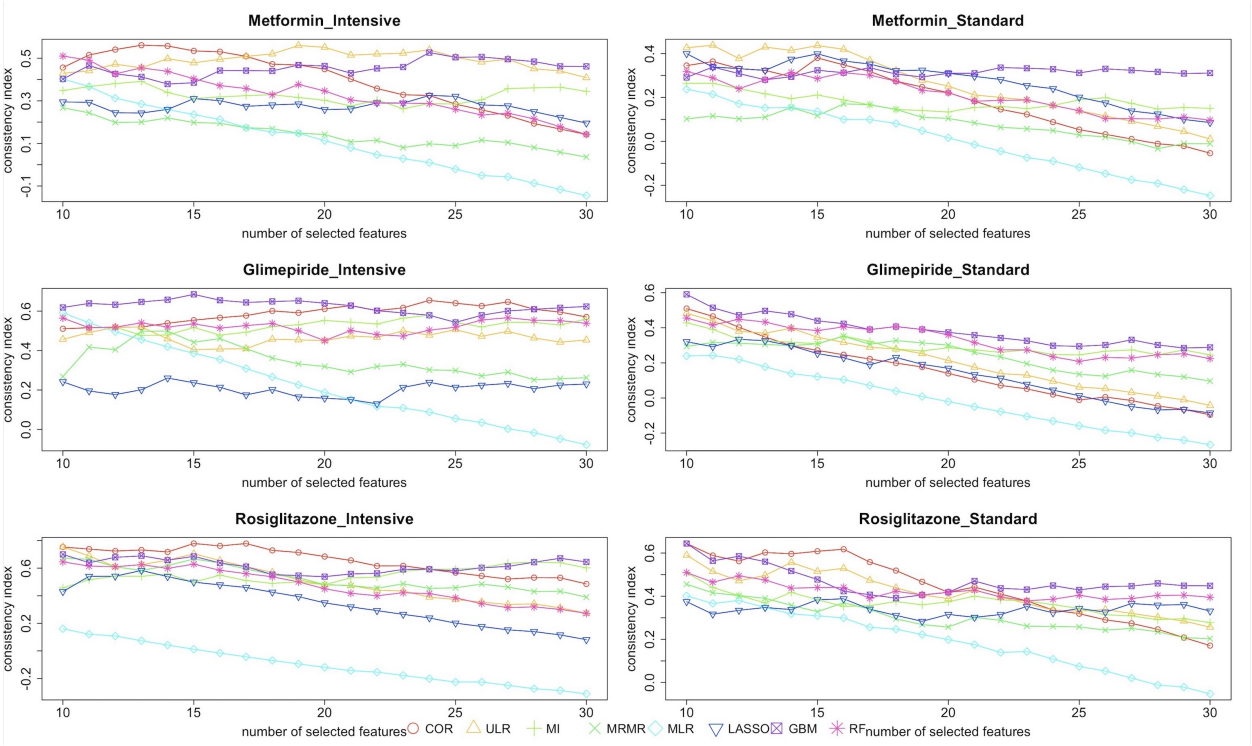

Figure 2: Consistency versus number of selected features for six cohorts

loss, waist circumference.

- For Rosiglitazone standard cohort: BMI, smoked cigarettes, DBP, eye disease, HbA1c at base, family history of heart disease; Aggregate score of sensation, mobility,cognition, self-care, emotion; Aggregate score of vision, hearing, speech, ambulation, dexterity, emotion, cognition and pain; FPG, SBP, visual acuity score of left eye, weight.

As the final feature sets show, the demographic features are the most common important features for all the six cohorts, indicating no matter what drug they take, their demographic information is always important to the HbA1c changes. For metformin cohort, intensive subgroup has many demographic features identified to be important compared with the standard subgroup, including race, BMI, furthermore, we can also find that most of the physiological states are important to intensive cohort, but very few for standard. For Glimepiride cohort, hypoglycemia status, HDL, triglyceride and some demographic features (waist circumference, weight) found to be predictive in the intensive cohort, but race, ALT play an important role in the standard treatment subgroup. For Rosiglitzone cohort, intensive subgroup are still more sensitive to physiological state as standard group is more prone to health score or family history, which is interesting.

Previous work can verify some of our discoveries[21], for example, Age, BMI, SBP, DBP, triglyceride, proteinuria are all in the top 15 predictive features. previous work[22] also indicates that Age plays an important role in risk of cardiovascular disease, intensive glucose lowering increased the risk of cardiovascular disease and total mortality in younger participants, whereas it had a neural effect in older participants. Bujac[23] proposed that fasting plasma glucose was significant, it[23] also applied meta analysis of nine studies to manage to confirm the effect of FPG. According to work[24], Age, BMI, and HGI(observed minus expected HbA1c derived from pre-randomization fasting plasma glucose) may help individualize prediction of the benefits and harm from intensive glycemic therapy. Tyler[25] also works on comparison between intensive and standard arm to check the factors associated with the level of HbA1c at the end point, younger age, female gender, higher BMI, longer duration of diabetes, higher baseline HbA1c, black race,

history of cardiovascular disease event(s) were associated with a 12-month HbA1c >= 8.0%, while factors related to failure to reach a 12-month HbA1c of <= 8.0% include: race, age, poorer baseline glucose control, insulin use, severe hypoglycemia and weight gain. Luo[26] suggests Age, diastolic blood pressure, high density lipoprotein, waist circumference, sex, cholesterol, parental or sibling history, BMI and triglyceride as set of important features.

5 Conclusion

In this paper, we systematically studied the impact of different feature selection algorithms to the predictive modeling problem.We specifically investigated the problem of response heterogeneity of popular T2DM drugs. Our results demonstrated that among all the algorithms we picked, GBM cannot only achieve good prediction performance, but also produce the most consistent feature set. We found a large overlap between the features picked by those approaches and the features identified in the literature from domain knowledge, which implies the clinical validity of the selected features.

6 Acknowledgement

This work is supported by NSF IIS-1716432 and IIS-1750326.

References

[1] Jonathan E Shaw, Richard A Sicree, and Paul Z Zimmet. Global estimates of the prevalence of diabetes for 2010 and 2030. *Diabetes research and clinical practice*, 87(1):4–14, 2010.

[2] World Health Organization, World Health Organization, et al. Diabetes fact sheet no. 312. *Geneva, Switzerland URL: http://www. who. int/mediacentre/factsheets/fs312/en/index.html*, 2011.

[3] Mozhgan Dorkhan and Anders Frid. A review of pioglitazone hcl and glimepiride in the treatment of type 2 diabetes. *Vascular health and risk management*, 3(5):721, 2007.

[4] Lilian Beatriz Aguayo Rojas and Marilia Brito Gomes. Metformin: an old but still the best treatment for type 2 diabetes. *Diabetology & metabolic syndrome*, 5(1):6, 2013.

[5] DREAM (Diabetes REduction Assessment with ramipril, rosiglitazone Medication) Trial Investigators, et al. Effect of rosiglitazone on the frequency of diabetes in patients with impaired glucose tolerance or impaired fasting glucose: a randomised controlled trial. *The Lancet*, 368(9541):1096–1105, 2006.

[6] Hussein Raef, Abdulraof Al-Mahfouz, and Abdullah Al-Khonaizan. Adding rosiglitazone to metformin in patients with type 2 diabetes: Effect on diabetes control and metabolic parameters. *International Journal of Diabetes Mellitus*, 1(1):2–6, 2009.

[7] Kristine Faerch, Adam Hulmán, and Thomas PJ Solomon. Heterogeneity of pre-diabetes and type 2 diabetes: implications for prediction, prevention and treatment responsiveness. *Current diabetes reviews*, 12(1):30–41, 2016.

[8] Ronald A Cantrell, Carlos I Alatorre, Elizabeth J Davis, Victoria Zarotsky, Elisabeth Le Nestour, G Cuyún Carter, Iris Goetz, Rosirene Paczkowski, and Justo Sierra-Johnson. A review of treatment response in type 2 diabetes: assessing the role of patient heterogeneity. *Diabetes, Obesity and Metabolism*, 12(10):845–857, 2010.

[9] Action to Control Cardiovascular Risk in Diabetes Study Group. Effects of intensive glucose lowering in type 2 diabetes. *New England journal of medicine*, 358(24):2545–2559, 2008.

[10] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003.

[11] Melissa J Azur, Elizabeth A Stuart, Constantine Frangakis, and Philip J Leaf. Multiple imputation by chained equations: what is it and how does it work? *International journal of methods in psychiatric research*, 20(1):40–49, 2011.

[12] Hanchuan Peng, Fuhui Long, and Chris Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis and machine intelligence*, 27(8):1226–1238, 2005.

[13] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[14] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[15] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[16] Mark W. Woolrich, Brian D. Ripley, Michael Brady, and Stephen M. Smith. Temporal autocorrelation in univariate linear modeling of fmri data. *NeuroImage*, 14(6):1370 – 1386, 2001.

[17] Andrew M. Fraser and Harry L. Swinney. Independent coordinates for strange attractors from mutual information. *Phys. Rev. A*, 33:1134–1140, Feb 1986.

[18] Frederick L. Oswald Nimon, Kim F. Understanding the results of multiple linear regression: Beyond standardized regression coefficients. *Organizational Research Methods*, 16:650674, 10 2013.

[19] T. Hastie J. Elith, J. R. Leathwick. A working guide to boosted regression trees. *Journal of Animal Ecology*, 77:802–813, 04 2008.

[20] P. Drotr, J. Gazda, and Z. Smkal. An experimental comparison of feature selection methods on two-class biomedical datasets. *Computers in Biology and Medicine*, 66:1 – 10, 2015.

[21] Norman Black Roy Harpe Yue Huang, Paul MaCullagh. Feature selection and classification model construction on type 2 diabetic patients data. *Artificial Intelligence in Medicine*, 47(3):251–262, 2007.

[22] Hertzel C. Gerstein Robert P. Byington William C. Cushman Henry N. Ginsberg Walter T. Ambrosius Laura Lovato William B. Applegate for the ACCORD Investigators Michael E. Miller, Jeff D. Williamson. Effects of randomization to intensive glucose control on adverse events, cardiovascular disease, and mortality in older versus younger adults in the accord trial. *Diabetes Care*, 37(3):634–643, 2014.

[23] Sarah Bujac, Angelo Del Parigi, Jennifer Sugg, Susan Grandy, Tom Liptrot, Martin Karpefors, Chris Chamberlain, and Anne-Marie Boothman. Patient characteristics are not associated with clinically important differential response to dapagliflozin: a staged analysis of phase 3 data. *Diabetes Therapy*, 5(2):471–482, 2014.

[24] Deborah J. Wexler Seth.A.Berkowitz Sanjay Basu, Sridharan Raghavan. Characteristics associated with decreased or increased mortality risk from glycemic therapy among patients with type 2 diabetes and high cardiovascular risk: Machine learning analysis of the accord trial. *Diabetes Care*, 41(3):604–612, 2018.

[25] Donald Hire S J Chen Robert M. Cohen R W Mcduffie Eric Sixtus Nylen Patrick J. O'Connor Saira Rehman Elizabeth R. Seaquist Tyler C Drake, F C Hsu. Factors associated with failure to achieve a glycated hemoglobin target of $<$8.0% in the action to control cardiovascular risk in diabetes (accord) trial. *Diabetes, obesity metabolism*, 18(1):92–5, 2016.

[26] Han Longfei Zeng Ping Chen Feng Pan Limin Wang Shu Zhang Tiemei Luo, Senlin. A risk assessment model for type 2 diabetes in chinese. *PLOS ONE*, 9(8):1–7, 08 2014.