DNDKDDR

October 2023

Abstract

Background: The study of diabetic microvascular complications represents a pivotal area of inquiry within the broader field of diabetes research. The identification of biomarkers for these complications is paramount for predicting their onset and monitoring their progression. It is evident that traditional biomarkers, such as HbA1c and Serum Creatinine (SCr), are limited in their ability to accurately determine complications under specific clinical conditions. Consequently, recent literature has increasingly focused on the discovery of non-traditional biomarkers.

Objective: The aim of this study is to identify novel potential biomarkers for two types of diabetic microvascular complications, diabetic kidney disease (DKD) and diabetic retinopathy (DR), from a large-scale observational dataset. Additionally, the study aims to assess the extent to which variations in these biomarkers influence the risk of developing such complications.

Methods: The majority of studies on biomarkers of diabetic microvascular complications are currently conducted through small randomized controlled trials. Nevertheless, these trials frequently necessitate the pre-specification of clinical indicators to be studied, involve cumbersome procedures, and have limited sample sizes.In order to overcome these obstacles, we selected features from a large-scale clinical observational dataset using machine learning approaches in an effort to identify biomarkers for microvascular complications. We then employed a combination of machine learning and causal inference methods to evaluate the effect of these biomarker variations on the likelihood that patients may have complications.

Conclusion: A clinical dataset comprising 1,984 diabetic patients was employed in this study. The majority of the biomarkers selected have been previously discussed in the literature. Furthermore, a novel biomarker for diabetic retinopathy (DR), total bilirubin (TBIL), and a novel biomarker for diabetic kidney disease (DKD), C-peptide, were identified. Causal inference assessment revealed that a patient's anticipated risk of developing DKD increases by 17% when their C-peptide level is below 1.35 ng/ml. Consequently, the probability of a patient developing DR is estimated to increase by 5% if their TBIL level is below 12.93 umol/L. These findings provide valuable insights that can inform future clinical studies. microvascular complications.

1 Introduction

The pervasive nature of unhealthy lifestyle habits has contributed to the high prevalence of type 2 diabetes mellitus (T2DM), leading to an increased incidence of microvascular complications. This situation imposes a heavy burden on social healthcare systems[1]. Among the prevalent microvascular complications associated with diabetes, diabetic kidney disease (DKD) and diabetic

retinopathy (DR) are significant concerns, as they can result in high mortality and blindness. Despite efforts to improve renal and ocular outcomes in diabetic patients, the proportion of DKD progressing to end-stage renal disease (ESRD) and DR to visual impairment remains persistently high globally[2,3]. Currently, there are no efficient methods for accurately detecting DKD and DR prior to the onset of classical symptoms. Therefore, a thorough exploration and enhanced comprehension of biochemical indicators could offer novel insights into the early identification of diabetic microvascular complications.

The current clinical diagnosis of DKD primarily relies on the assessment of urinary microalbuminuria to urine creatinine ratio (UACR) and glomerular filtration rate (eGFR), while DR diagnosis is based on retinal imaging. However, early diagnosis of DKD and DR is limited if only these indicators or methods are used. Efforts have been made to improve diagnostic capabilities by exploring additional predictors. A systematic review and meta-analysis recently indicated that common predictors of DKD and DR include age, sex, glycated hemoglobin (HbA1c), duration of diabetes, systolic blood pressure (SBP), and body mass index (BMI)[4]. Recent studies have also identified other clinical indicators for predicting DKD, such as uric acid (UA), soluble tumor necrosis factor receptor (TNFR), and lipid profiles[5,6,7]. In the case of DR, novel indicators such as anemia, dyslipidemia, and microalbuminuria have been recognized[8,9].

Furthermore, several studies have focused on the combination of markers to improve prediction accuracy. For instance, Yun and Bilgin revealed significant associations between the triglyceride/high-density lipoprotein cholesterol (TG/HDL-C) ratio and the development of DKD, as well as the C-reactive protein to serum albumin ratio (CAR) and DKD[10,11]. Additionally, accumulating evidence demonstrates that DR is a strong predictor of DKD, and vice versa[12]. The interaction between DKD and DR complicates the diagnosis of microvascular complications in diabetic patients.

To better identify common and distinctive predictors for DKD and DR, this study will focus on the following aspects in predicting microvascular complications associated with diabetes:

1. Utilizing machine learning methods for feature selection to identify the most suitable set of features for predicting DKD/DR.

2. Comparing predictors between DKD and DR.

3. Establishing a causal inference model to evaluate the relationship between predictors and DKD/DR.

This research aims to identify more representative indicators to detect high risks of DKD or DR and to enhance comprehensive knowledge of their pathogenesis.

2 Methodology and Materials

The objective of our research is to identify clinical markers associated with microvascular complications in diabetes and to evaluate how variations in these markers influence the risk of developing such complications. To accomplish this objective, we utilized a hybrid feature selection methodology that combines filter and wrapper methods to identify clinical markers associated with the respective complications. Subsequently, to evaluate the impact of changes in these clinical biomarkers on disease risk, we developed a causal inference model based on physician recommendations and existing literature. In contrast to conventional statistical inference methods, the use of causal inference

algorithms facilitates the control of confounding factors, providing more accurate estimates of risk changes. Additionally, this approach enables counterfactual inference (e.g., estimating how the risk of complications would change if all patients had blood pressure levels above or below a certain threshold).

The study utilized patient data from the People's Hospital of Shanxi Province, specifically focusing on patients with diabetic microvascular complications. The dataset comprised 1,984 type 2 diabetic patients, among whom 351 had Diabetic Retinopathy (DR), 355 had Diabetic Kidney Disease (DKD), and 1,278 were type 2 diabetic patients without complications. The objective of our study is to separately investigate the clinical biomarkers influencing diabetic kidney disease (DKD) and diabetic retinopathy (DR). In order to achieve this objective, the dataset was preprocessed and divided into two sub-datasets. In the DKD sub-dataset, patients with DKD were labeled as 1, while patients with type 2 diabetes without complications were labeled as 0. Similarly, in the DR subset, patients with DR were labeled as 1, and patients without complications were labeled as 0. Figure 1 shows the results of splitting the dataset.

Could we also have a chart for patient selection?

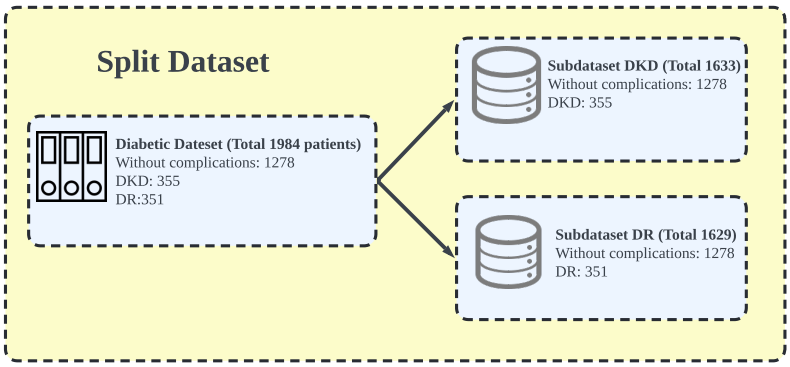


Figure 1: Split Dataset for Analysis: Dividing the complete dataset into DKD and DR datasets for separate investigation of different complications.

The study focused on two specific complications: Diabetic Kidney Disease (DKD) and Diabetic Retinopathy (DR). We identified distinct sets of clinical biomarkers associated with each complication and compared the similarities and differences between these sets. Subsequently, the study evaluated the impact of alterations in several clinical biomarkers on the risk of developing the corresponding complications. Figure 2 illustrates the entire workflow of the study analysis

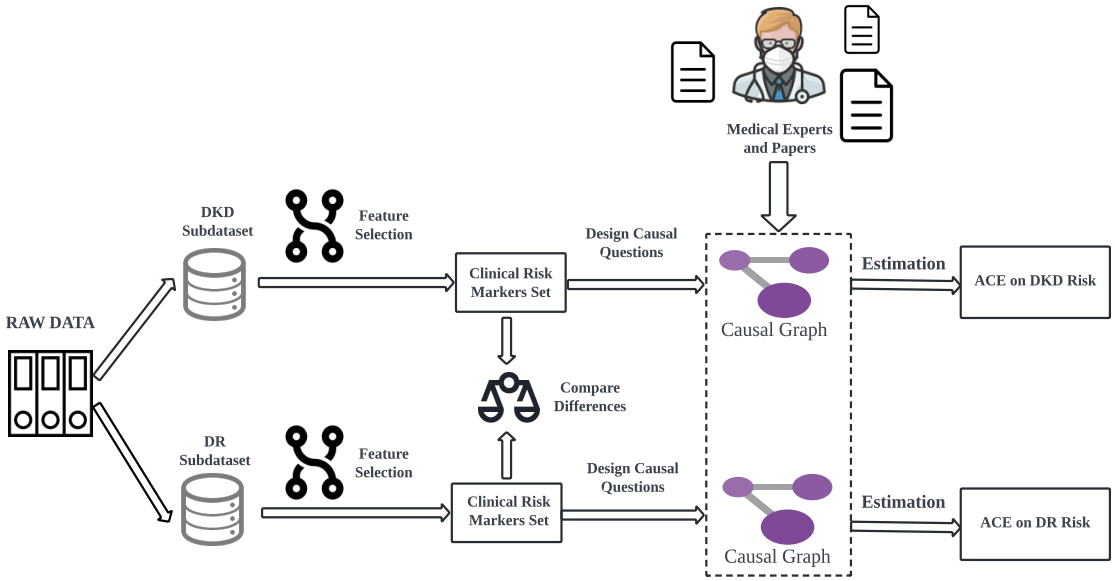


Figure 2: Entire Workflow of the Study Analysis: Separately investigate two types of complications and compare differences in feature sets after feature selection.

2.1 Hybrid Feature Selection

In this study, we propose a hybrid feature selection model to ensure that the selected features have a strong correlation with the complications and significantly contribute to predicting the occurrence of these complications. The hybrid feature selection process comprises two principal stages: a filter stage and a wrapper stage. Figure 3 depicts the complete workflow of the hybrid feature selection process. The filter stage encompasses correlation tests and LASSO regression. Although LASSO regression is typically regarded as an embedded feature selection method, it is typically employed to describe linear relationships and has limited predictive power for nonlinear relationships (e.g., when no nonlinear terms are assumed). [8] Consequently, in this context, it is employed solely as a filter, rather than as a final predictive model. The wrapper stage entails the selection of variables based on the performance of predictive models. The choice of XGBoost and Random Forest as predictive models was motivated by their demonstrated capacity to effectively capture linear and nonlinear relationships. Furthermore, both models are ensemble methods, which can effectively enhance model performance.[4] Moreover, these models can output feature split thresholds in decision trees, each split node in a decision tree contains a certain amount of information, which provide valuable references for the discretization of clinical indicator changes in subsequent causal inference analysis. [20] To ensure the reliability of the information at these split nodes, we consulted with physicians after calculating the split thresholds. By incorporating the physicians' opinions, we confirmed which variables could be split at which thresholds.

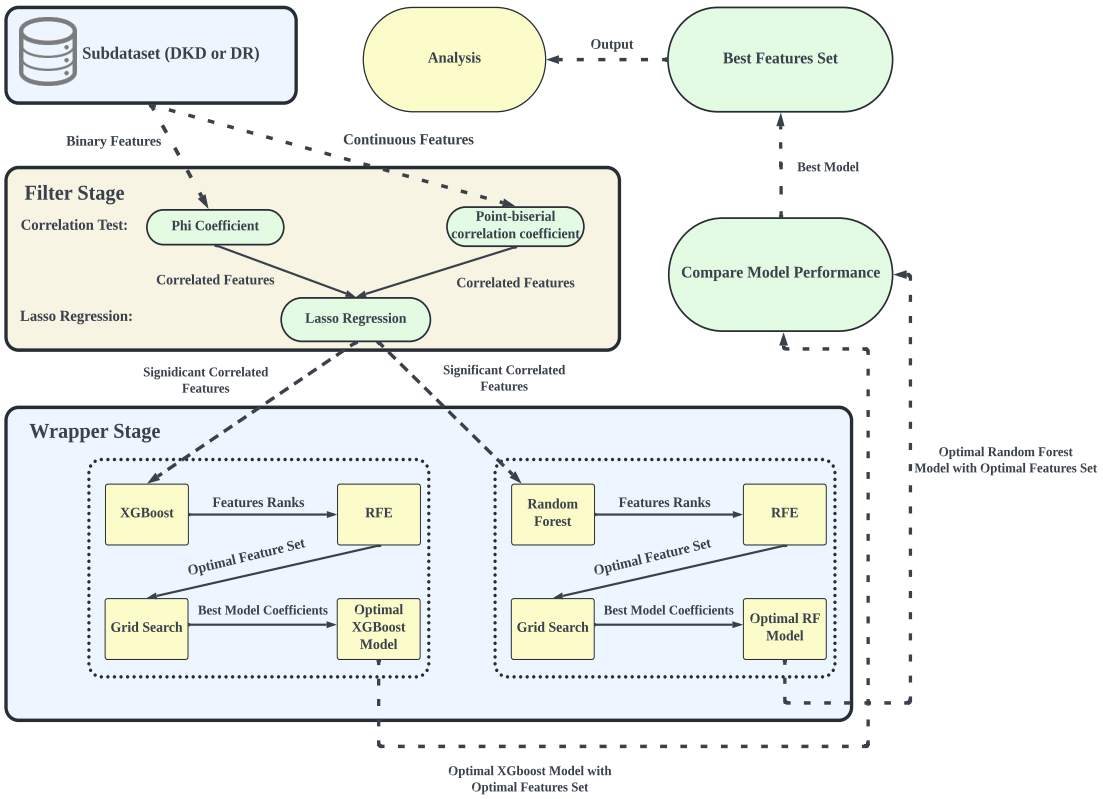


Figure 3: Flow of Hybrid Feature Selection (HFS): After filtering out insignificantly correlated variables, two prediction models were constructed within the wrapper, and finally, the optimal feature set was selected by comparing the performance of different prediction models.

Figure 3 illustrates the detailed steps of the Hybrid Feature Selection process.

After preprocessing the data, we input a sub-dataset corresponding to a particular complication and denote all clinical markers within this subset as the feature set $\mathcal{F}_{Total}$.

(1): First, correlation tests are used to preliminarily identify which features are correlated to the complication. Since our complication is a binary variable (0 indicates no complication, 1 indicates the presence of a complication) and the features include both binary and continuous variables, we use different correlation tests to handle the different data types appropriately.

The point-biserial correlation coefficient [5] is employed in correlation analysis between continuous features and the complication. For binary features and the complication, we employed the Phi Coefficient.[7] After performing the correlation test, we obtain a feature set $\mathcal{F}_{Corr}$

(2) Secondly, we utilized the feature set $\mathcal{F}_{Corr}$ to construct a Lasso regression model. By transforming the patient features data (features in $\mathcal{F}_{Corr}$) into a matrix, we can obtain an np matrix X where n represents the number of patients and p represents the number of features in the

$\mathcal{F}_{Corr}$. Then, the target function of Lasso in terms of this formula [22]:

$$\min_{\beta} \frac{1}{2n} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1$$

β is a coefficient vector of length p, Lasso regression eliminates non-significant clinical markers by shrinking some feature coefficients to zero.[22] λ is the regularization parameter. To obtain the optimal regularization parameter λ_{opt}, we employ cross-validation and utilize the Area Under the Curve (AUC) as the selection criterion.[11] Thereafter, λ_{opt} is used to construct a LASSO regression model, which then filters out certain features, thus resulting in a feature set, designated as $\mathcal{F}_{sig}$.

(3) In the third step, predictive models are constructed to predict the occurrence of complications using the feature set $\mathcal{F}_{sig}$. The XGBoost model and Random Forest models are chosen for this purpose. The XGBoost model $\mathcal{M}_{XG}$ can identify the importance of features by evaluating the gain or coverage when splitting tree nodes. [6] The features in $\mathcal{F}_{sig}$ are then sorted in ascending order based on their importance, denoted as R_{XG}. Random Forest model $\mathcal{M}_{RF}$ identify the importance of features by internal out-of-bag estimation, and sort $\mathcal{F}_{sig}$ features in ascending order based on feature importance, denotes as R_{RF}.[4]

(4) We subsequently employ recursive feature elimination (RFE) to conduct feature selection [9]. The Recursive Feature Elimination (RFE) process can be represented by the following equation:

$$\text{RFE}(\mathcal{M}, \text{step=1}) = \arg\max_{\mathcal{F}_i \subset R} \text{AUC}_{\text{fold}} \left(\mathcal{M}, \mathcal{F}_i \setminus \{R_i\}\right)$$

$\mathcal{M}$ represents the predictive model (either $\mathcal{M}_{XG}$ or $\mathcal{M}_{RF}$), and "step = 1" indicates that one variable is removed from the model per iteration. R denotes the ordered set of features (R_{XG} or R_{RF}) for Random Forest), with $\mathcal{F}_i$ being a subset of R, representing the remaining features in the model after the i-th iteration. R_i represents the i-th feature in the ordered set R. Then, $\mathcal{F}_i/R_i$ means that, at each iteration, the feature with the smallest importance is removed from the model. The process can be represented by the following equation: in each iteration, the least important feature is removed from the model according to the feature importance ranking. Cross-validation is then used to calculate the AUC.[11] The model with the highest AUC is ultimately selected.

The model with the highest AUC is then subjected to a grid search in order to identify the optimal parameters. This procedure results in an optimal XGBoost model, denoted $\mathcal{BM}_{XG}$, and the feature set used in this model, denoted $\mathcal{F}_{XG}$. Similarly, an optimal random forest model, denoted as $\mathcal{BM}_{XG}$, is obtained along with its corresponding feature set, denoted as $\mathcal{F}_{XG}$.

(5) In the final step, the two models are compared by evaluating their Accuracy and Area Under the Curve (AUC) metrics. The model exhibiting the optimal performance in these metrics is designated as the optimal model, designated as $\mathcal{BM}$, along with its corresponding feature set, designated as $\mathcal{F}_{final}$.

If the DKD data subset is employed, the $\mathcal{F}_{final}$ is redefined as $\mathcal{F}_{DKD}$. Conversely, in the case of the DR data subset, $\mathcal{F}_{final}$ is redefined as $\mathcal{F}_{DR}$.This allows us to compare $\mathcal{F}_{DKD}$ and $\mathcal{F}_{DR}$, identifying overlapping and non-overlapping features. Such comparisons may provide valuable medical insights.

2.2 Causal Inference with Confounders

During the feature selection process, we identified several clinical markers that are indicative of either DKD or DR. Our next objective is to delve deeper into understanding how variations in

clinical markers influence the incidence of the complication. Based on existing research, we selected one feature each from $\mathcal{F}_{DKD}$ and $\mathcal{F}_{DR}$ for further study. We binarized these two features using the split thresholds derived from the final predictive model and incorporating physician recommendations. The central question we aim to address is: if all patients had values for a specific feature above the threshold compared to all patients having values below the threshold, what would be the resultant difference in the risk of developing the complication? In order to address this question, we proposed a causal inference model.

However, the relationship between the clinical marker and the complication was influenced by several confounders. To maintain the accuracy of the causal inference model, it's imperative to address confounders. Following physician advice, we classified these confounders into two categories: The first category of confounders were objective signs inherent to the patient, such as age, sex, and BMI, which usually remained unaffected by other factors but had an impact on the development of the complication and the studied clinical marker. Another category of confounders were clinical markers that correlated with the biomarker we wanted to study. However, the observed correlation is a result of changes in the functioning of a specific body organ or part, leading to alterations in both the biomarker we want to study and the other biomarkers. To obtain a more precise estimation of the impact of a biomarker on complications, we created a Directed Acyclic Graph (DAG) for a causal inference model, taking into account the confounding factors mentioned earlier. What does "DAG" stand for?

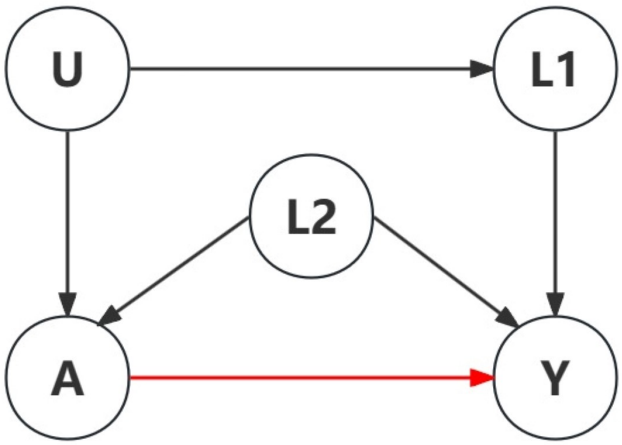


Figure 4: DAG of Causal Inference: Describes the causal relationships between intervention, outcome, and confounding factors.

This Figure 4 depicts the studied clinical marker, denoted by A, which is discretized into a binary variable (0 or 1). A value of 1 can be seen as an intervention, while a value of 0 indicates no intervention. The outcome of interest is a diabetic microvascular complication (DKD or DR), denoted by Y. Then, the effect of changes in the studied biomarker on the complication of interest is the causal effect of A on Y. $L1$ represents objective patient characteristics, such as age, sex, and BMI. $L2$ represents some other biomarkers of the patient, which are usually indicators of certain organs in the body, reflecting the function of certain organs or changes in the body. These

biomarkers also have an impact on complications. U represents changes in organ function or bodily changes that are difficult to measure.

In this causal diagram, it is necessary to assume that there are no additional unmeasured common causes between A and Y, apart from the listed factors. There are two backdoor paths between A and Y: A←L1→Y and A←U→L2→Y. Both of these paths can cause bias when estimating the effect of A on Y. Therefore, it is essential to block these two backdoor paths. According to the backdoor criterion, controlling $L1$ and $L2$ can block all backdoor paths, allowing for an accurate estimation of the effect of A on Y, as U is typically difficult to measure.

Then, we need to satisfy several assumptions: First, we must ensure consistency. This means that we regard A as the intervention, our observed intervention is $A = a$ (a=0,1), and the observed outcome is Y. The counterfactual outcome is denoted as Y^a. The observed outcome Y under the condition $A = a$ is consistent with the counterfactual outcome Y^a. Second, we must ensure conditional exchangeability. Under the condition of blocking all backdoor paths from A to Y, the counterfactual outcome Y^a is independent of the observed intervention. This means that Y^a is independent of A, conditional on $L1$ and $L2$ ($Y^a \perp A \,|L1, L2$). Additionally, it is important to ensure positivity by controlling for the conditions L1 and L2. This means that the probability of each individual receiving the intervention or not receiving the intervention is greater than 0 ($P(A = a|L1,L2) > 0$).

In order to assess causal effects, we elected to utilise two metrics: the Risk Difference (RD) and the Risk Ratio (RR). The RD represents the difference in the probability of disease between the intervention group and the control group, while the RR represents the ratio of the probability of disease between the intervention group and the control group. In the field of causal inference, RD is often referred to as the Average Causal Effect (ACE). Since our outcome Y is a binary variable (indicating the presence or absence of complications), RD and RR can be expressed using the following formula:

$$\widehat{RD} = \widehat{ACE} = P(Y^{a=1}|L1, L2) - P(Y^{a=0}|L1, L2)$$

$$\widehat{RR} = \frac{P(Y^{a=1}|L1, L2)}{P(Y^{a=0}|L1, L2)}$$

When all backdoor paths from A to Y are blocked, the ACE of A on Y that we wish to investigate can be expressed as:

$$\widehat{ACE} = P(Y^{a=1}|L1, L2) - P(Y^{a=0}|L1, L2)$$

Due to conditional exchangeability, the equation can be transformed into:

$$\widehat{ACE} = P(Y^{a=1}|A = 1, L1, L2) - P(Y^{a=0}|A = 0, L1, L2)$$

Due to consistency:

$$\widehat{ACE} = P(Y|A = 1, L1, L2) - P(Y|A = 0, L1, L2)$$

This process converts unobservable counterfactual outcomes $P(Y^{a=1}|L1, L2)$ and $P(Y^{a=0}|L1, L2)$ and into observable outcomes $P(Y|A = 1, L1, L2)$ and $P(Y|A = 0, L1, L2)$, allowing us to calculate the Average Causal Effect (ACE) of A on Y based on the available data.

Correspondingly, RR can also be transformed into an observable outcome, and it can be expressed using the following formula:

$$\widehat{RR} = \frac{P(Y|A=1, L1, L2)}{P(Y|A=0, L1, L2)}$$

The problem remaining is how to control $L1$ and $L2$. The Augmented Inverse Probability Weighted (AIPW) method is primarily utilized for this purpose. The main concept behind AIPW is to generate virtual samples by assigning individual weights to eliminate the impact of confounding factors. AIPW is a doubly robust method, which typically results in smaller errors and greater stability compared to other weighting methods such as IPW (Inverse Probability Weighting). We validated the results obtained through AIPW by employing the Fast Large-scale Almost Matching Exactly Approach (FLAME) method to calculate average causal effects. FLAME uses a matching approach, where each patient in the treatment group is matched with a similar patient in the control group based on their characteristics.[24] Therefore, the estimated average causal effect is the overall difference in complication risks between patients in the treatment and control groups.

Because these two methods use different principles to calculate the Average Causal Effect (ACE), achieving similar results in their computations indicates a high level of credibility and robustness in the ACE that has been calculated.

2.2.1 Augmented Inverse Probability Weighting

The Augmented Inverse Probability Weighted (AIPW) estimator is a method for estimating average treatment effects that combines properties of data-adaptive estimation (e.g., Super Learning) and inverse probability weighted estimators, making it a "doubly robust" method. The AIPW estimator is called 'doubly robust' because it is consistent as long as either the treatment assignment mechanism or the outcome model is correctly specified. This property allows the AIPW estimator to provide more reliable results in situations where the treatment assignment process and outcome model are uncertain. [19] [13]

The Augmented Inverse Probability Weighting (AIPW) estimator estimates the Average Causal Effect (ACE) by following two basic steps: first, fitting a propensity score model to estimate the probability of different interventions ($P(A|L)$) based on observed covariates ($L = L1, L2$). This probability is called the propensity score, we denote it as $\widehat{e}(L) = P(A|L)$. And second, fitting two models to estimate the outcome under intervention ($A = 1$) and non-intervention ($A = 0$) conditions. The XGBoost model was used in this step to predict the outcomes of two different interventions. The results of this step are represented as $\widehat{m}(1, L) = P(Y|A = 1, L)$ and $\widehat{m}(0, L) = P(Y|A = 0, L)$. Finally, the ACE estimate is calculated using the provided formula. The steps for calculating it are as follows:

Algorithm 1 Augmented Inverse Propensity Weighting (AIPW) for $\widehat{\text{ACE}}_{\text{AIPW}}$

Require:
Observational data: (Y, A, L)
Propensity score model: $\hat{e}(L)$

Ensure:
Causal effect estimate: $\widehat{\text{ACE}}_{\text{AIPW}}$

1: **Step 1:** *Estimate Propensity Scores* [Logistic Regression of treatment on covariates]
$\hat{e}(L) \leftarrow$ Fit a logistic regression model on $\{A \sim L\}$

2: **Step 2:** *Estimate Outcome Models* [Prediction of outcome on treatment and covariates]
$\hat{m}(1, L) \leftarrow$ Fit a XGboost model on $\{Y \sim A + L\}$
$\hat{m}(0, L) \leftarrow$ Fit a XGboost model on $\{Y \sim A + L\}$

3: **Step 3:** *Compute AIPW Estimate for* $\widehat{\text{ACE}}_{\text{AIPW}}$
$\widehat{\text{ACE}}_{\text{AIPW}} = \frac{1}{n}\sum_{i=1}^{n}((\hat{m}(1, L_i) - \hat{m}(0, L_i)) + (\frac{A_i(Y_i - \hat{m}(1,L_i))}{\hat{e}(L_i)} - \frac{(1-A_i)(Y_i - \hat{m}(0,L_i))}{1-\hat{e}(L_i)}))$

4: **Output:** $\widehat{\text{ACE}}_{\text{AIPW}}$

Because the AIPW estimator only requires one of the propensity or outcome models to be correctly specified, unlike other methods that require both models to be accurate. This double robustness property makes the AIPW estimator less biased and more reliable in estimating average treatment effects, especially in scenarios where one of the models may be misspecified.[18]

2.2.2 Fast Large-scale Almost Matching Exactly Approach

FLAME (Fast Large-scale Almost Matching Exactly) is a method proposed for high-quality almost-exact matching in high-dimensional categorical datasets for estimating conditional average treatment effects (CATEs).[24] FLAME leverages a hold-out training set to learn a distance metric for matching and progressively drops covariates to achieve high-quality matches while maintaining balance between treatment and control groups.[24]

FLAME has several key advantages, including its ability to scale to large datasets with millions of observations where other methods fail. It also achieves significantly better performance than other matching techniques. FLAME achieves significantly better performance than coarsened exact matching, mixed integer programming matching, and network flow methods by learning the distance metric instead of introducing it a priori. This means that FLAME does not suffer when irrelevant variables are introduced.[2][12]

3 Result

3.1 Hybrid Feature Selection

Table 1 and Table 2 present the 38 variables in our dataset. Table 1 comprises 32 clinical markers, encompassing various hematologic parameters, alongside demographic factors such as age, BMI, smoking status, and alcohol consumption. Table 2 delineates medication usage among patients, encompassing a total of 120 distinct drugs, categorized into six classes: ACE inhibitors (ACEI), angiotensin II receptor blockers (ARB), beta-blockers, calcium channel blockers (CCB), diuretics, insulin, and statins.

Table 1: Table of Abbreviations and Full Names for Clinical Biomarkers

Abbreviation	Full name	Unit	Data Type
Age	Patient Age	Year	Integer
ALB	Albumin	g/L	Continous
Alcohol	Alcohol consumption status (0: Never Consumed, 1: Consumed)	0 or 1	Binary
ALT	Alanine aminotransferase	IU/L	Continous
AST	Aspartate aminotransferase	IU/L	Continous
BMI	Patient Body Mass Index	Kg/mˆ2	Continous
BUN	Urea nitrogen	mmol/L	Continous
CHO	Total cholesterol	mmol/L	Continous
C-peptide	Fasting C-peptide	ng/ml	Continous
DBIL	Direct bilirubin	umol/L	Continous
DBP	Diastolic Blood Pressure	mmHg	Continous
Gender	Patient Gender (0:Female, 1:Male)	0 or 1	Binary
GLUC	Urine Glucose	mmol/L	Continous
HbAlc	Glycated hemoglobin	%	Continous
HCT	Hematocrit	%	Continous
HDL	High-density lipoprotein cholesterol	mmol/L	Continous
IBIL	Indirect bilirubin	umol/L	Continous
LDL-C	Low-density lipoprotein cholesterol	mmol/L	Continous
Na	Sodium	mmol/L	Continous
PLT	Platelet count	$\times$10ˆ9/L	Continous
RBC	Red blood cells	$\times$10ˆ12/L	Continous
RDWCV	Variation coefficient of red blood cell distribution width	%	Continous
RDWSD	Standard deviation of red blood cell distribution width	fL	Continous
SBP	Systolic Blood Pressure	mmHg	Continous
SCr	Serum Creatinine	umol/L	Continous
Smoke	Patient smoking status (0: Never Smoked, 1: Smoked)	0 or 1	Binary
TBIL	Total serum bilirubin	umol/L	Continous
TG	Triglycerides	mmol/L	Continous
TSH	Thyroid stimulating hormone	uIU/ml	Continous
UA	Uric acid	umol/L	Continous
WBC	White blood cell count	$\times$10ˆ9/L	Continous

Table 2: Table of Drug Abbreviations and Full Names: "0" indicates the drug was not used, while "1" indicates the drug was used during the treatment of diabetes.

Abbreviation	Full name	Unit	Data Type
ACEI	Angiotensin-converting enzyme inhibitor usage status	0 or 1	Binary
ARB	Angiotensin II receptor blocker usage status	0 or 1	Binary
BetaBlockers	BetaBlockers usage status	0 or 1	Binary
CCB	Calcium channel blocker usage status	0 or 1	Binary
Diuretic	Diuretic usage status	0 or 1	Binary
Insulin	Insulin usage status	0 or 1	Binary
Statins	Statins usage status	0 or 1	Binary

The present study investigated two complications: diabetic kidney disease (DKD) and diabetic retinopathy (DR). Following the hybrid feature selection process, the indicative factors associated

with DKD and DR were identified and are detailed in Tables 3 and 4, respectively. In Tables 3 and 4, the designation "Normal" refers to the control group patients, who represent the typical type 2 diabetes patient population. These "Normal" patients have diabetes but neither DKD nor DR. The indicative factors that we have identified include both continuous and binary variables, and we have separated them accordingly in the statistical tables. For continuous variables, we calculated the mean and confidence intervals. For binary variables, the mean proportion and confidence intervals for the category coded as "1" were calculated. Due to the non-normal distribution of our data, we employed the Mann-Whitney U (MW) test to detect whether there are differences in the means of continuous variables between "Normal" patients and those with complications. Similarly, the Chi-Square (χ^2) test was employed to assess the disparity in proportion means between the two groups for binary variables.

Table 3: Table of Indicative Factors for DKD

Indicative Factors	**Mean Value**		**Standard Deviation**		**MW Test**
Continous	**Normal**	**DKD**	**Normal**	**DKD**	**P-value**
ALB (g/L)	41.4 ± 0.21	40.46 ± 0.5	3.8	4.76	2.76E-03
AST (IU/L)	20.25 ± 0.61	22.76 ± 1.84	11.13	17.64	1.49E-02
BMI (Kg/mˆ2)	25.31 ± 0.17	26.77 ± 0.76	3.18	7.24	3.74E-07
BUN (mmol/L)	5.38 ± 0.08	5.99 ± 0.23	1.43	2.18	1.11E-06
C-peptide (ng/ml)	1.56 ± 0.04	1.07 ± 0.07	0.76	0.66	8.83E-35
DBP (mmHg)	78.07 ± 0.45	81.26 ± 1.2	8.27	11.49	4.56E-08
GLUC (mmol/L)	8.39 ± 0.18	9.46 ± 0.4	3.2	3.79	6.80E-08
HbAlc (%)	8.69 ± 0.09	9.06 ± 0.2	1.58	1.88	1.41E-03
LDL-C (mmol/L)	2.71 ± 0.04	2.86 ± 0.09	0.75	0.85	4.06E-03
PLT (×10ˆ9/L)	199.76 ± 3.21	208.37 ± 7.21	58.43	68.98	4.73E-02
RBC (×10ˆ12/L)	7.72 ± 3.13	6.85 ± 2.03	57.12	19.39	4.48E-09
SBP (mmHg)	127.62 ± 0.69	133.93 ± 1.7	12.58	16.25	7.34E-11
SCr (umol/L)	67.66 ± 0.96	76.45 ± 2.73	17.47	26.12	1.45E-09
TG (mmol/L)	2.1 ± 0.11	2.7 ± 0.28	1.93	2.69	4.40E-07
UA (umol/L)	306.48 ± 4.53	340.75 ± 10.23	82.46	97.86	7.50E-09
Indicative Factors	**Mean Proportion**		**Standard Deviation**		χ^2 **Test**
Binary	**Normal**	**DKD**	**Normal**	**DKD**	**P-value**
ACEI	0.01 ± 0.00	0.05 ± 0.02	0.08	0.21	6.06E-08
Alcohol	0.03 ± 0.01	0.24 ± 0.04	0.16	0.43	7.20E-42
ARB	0.06 ± 0.01	0.2 ± 0.04	0.24	0.4	5.48E-16
CCB	0.11 ± 0.02	0.36 ± 0.05	0.32	0.48	1.28E-28
Insulin	0.16 ± 0.02	0.37 ± 0.05	0.36	0.48	4.67E-18
Smoke	0.11 ± 0.02	0.34 ± 0.05	0.31	0.47	4.30E-25

Table 4: Table of Indicative Factors for DR

Indicative Factors	**Mean Value**		**Standard Deviation**		**MW Test**
Continous	**Normal**	**DR**	**Normal**	**DR**	**P-value**
ALB (g/L)	41.4 ± 0.21	40.87 ± 0.4	3.8	3.81	2.73E-02
BMI (Kg/mˆ2)	25.31 ± 0.17	25.77 ± 1.9	3.18	18.1	5.89E-03
C-peptide (ng/ml)	1.56 ± 0.04	1.29 ± 0.06	0.76	0.56	2.43E-09
DBIL (umol/L)	2.78 ± 0.27	2.41 ± 0.11	4.95	1.04	2.14E-02
HCT (%)	0.43 ± 0.00	0.42 ± 0.01	0.04	0.05	4.33E-03
IBIL (umol/L)	12.14 ± 0.33	11.15 ± 0.46	6.05	4.34	2.74E-03
SBP (mmHg)	127.62 ± 0.69	130.19 ± 1.48	12.58	14.12	2.44E-03
SCr (umol/L)	67.66 ± 0.96	67.93 ± 2.95	17.47	28.08	3.70E-02
TBIL (umol/L)	14.93 ± 0.56	13.6 ± 0.55	10.23	5.24	7.39E-03
Indicative Factors	**Mean Proportion**		**Standard Deviation**		χ^2 **Test**
Binary	**Normal**	**DR**	**Normal**	**DR**	**P-value**
Insulin	0.16 ± 0.02	0.2 ± 0.04	0.36	0.4	5.01E-02

Table 3 presents 21 indicative markers associated with diabetic kidney disease (DKD). All p-values derived from the MW test and chi-square test are less than 0.05, indicating that the observed differences between the control group and the complication group patients are highly significant for the indicative markers selected through Hybrid Feature Selection (HFS). Among these factors, ALB, C-peptide, and RBC exhibit a negative correlation with the risk of DKD, while the remaining 18 factors show a positive correlation with the risk of DKD. Significant discrepancies were observed in the mean values of C-peptide and uric acid (UA). The mean fasting C-peptide level for typical diabetes patients was 1.56 ng/ml, while the mean uric acid level was 306.48 umol/L. In contrast, for patients with diabetic kidney disease (DKD), the mean fasting C-peptide level is 1.07 ng/ml, while the mean uric acid level is 340.75 umol/L. In addition, traditional DKD markers such as SCr, HbA1c, and the proportions of smoking and alcohol consumption, also exhibit highly significant differences between "Normal" diabetes patients and those with DKD. Another noteworthy observation is the proportion of patients using calcium channel blockers (CCBs), angiotensin receptor blockers (ARBs), and insulin. For patients with normal diabetes, the mean proportion of patients using ARBs is approximately 6%, for CCBs it is 11%, and for insulin it is 16%. In contrast, for patients with DKD, the proportion of patients using ARB is approximately 20%, for CCB it is 36%, and for insulin it is 37%.

Table 4 presents 10 indicative factors associated with DR. Among them, ALB, C-peptide, DBIL, HCT, IBIL, and TBIL are negatively correlated with the risk of DR, while BMI, SBP, SCr, and the proportion of insulin use are positively correlated with the risk of DR. It is noteworthy that, in addition to the traditional DR biomarker SCr, the levels of bilirubin-related indicators TBIL, DBIL, and IBIL are generally lower in DR patients compared to normal diabetes patients. The mean TBIL level in DR patients is 13.6 umol/L, IBIL is 11.15 umol/L, and DBIL is 2.41 umol/L. In contrast, in normal diabetes patients, the mean TBIL level is approximately 15 umol/L, IBIL is 12.14 umol/L, and DBIL is 2.78 umol/L.

A comparison of the indicative factors in Tables 3 and 4 reveals that the shared indicative factors between the two tables are consistently associated with the complications, though the association is stronger with DKD. For example, SCr is positively correlated with both the risk of DKD and DR. However, the difference in SCr levels between DKD patients and "normal" diabetes patients

is considerably greater than that between DR patients and "normal" diabetes patients. The same conclusion can be drawn for factors such as ALB, BMI, C-peptide, SBP, and the proportion of insulin use. Moreover, in comparison to DR, DKD exhibits a considerably stronger correlation with blood pressure indicators and antihypertensive medications. A significantly higher proportion of patients with DKD utilize CCB and ARB medications compared to those with "Normal" diabetes. Furthermore, the differences in SBP and DBP between the two groups are highly significant. In contrast, patients with diabetic retinopathy (DR) are only associated with the SBP factor, and the difference is relatively smaller. In contrast to DKD, DR is significantly associated with bilirubin indicators TBIL, DBIL, and IBIL. This finding may provide insights into the pathological mechanisms of DR and catalyze further research in this area.

Table 5: Performance of Various DKD Prediction Models with and without Hybrid Feature Selection (HFS)

DKD Prediction Model	**Accuracy**	**Precision**	**F1 Score**	**Recall**
XGboost	0.856	0.827	0.647	0.531
Random Forests	0.865	0.712	0.649	0.597
Elastic Net	0.810	0.721	0.500	0.383
SVM	0.823	0.735	0.554	0.444
KNN	0.774	0.421	0.320	0.258
XGboost+HFS	**0.902**	0.826	**0.704**	**0.613**
Random Forests+HFS	0.878	**0.894**	0.656	0.519
Elastic Net+HFS	0.862	0.681	0.587	0.516
SVM+HFS	0.883	0.810	0.667	0.567
KNN+HFS	0.792	0.636	0.275	0.175

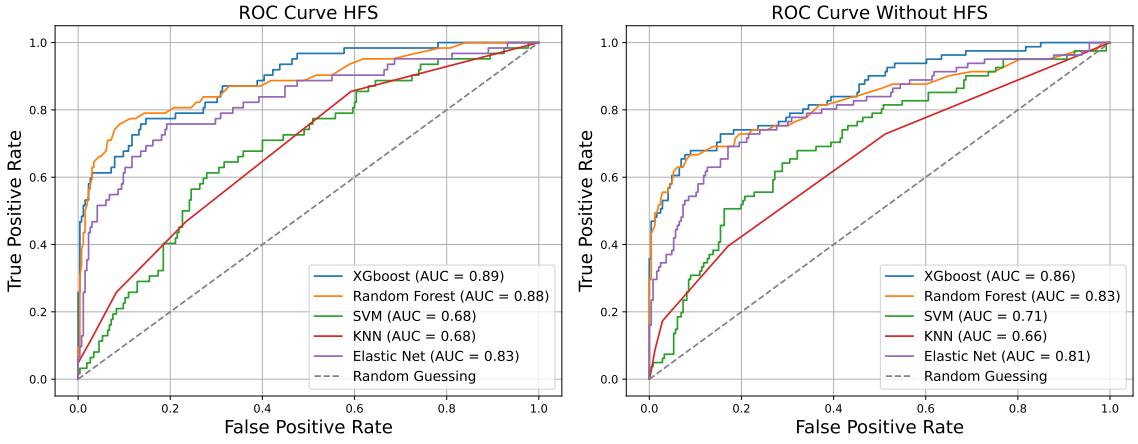
Table 6: Performance of Various DR Prediction Models with and without Hybrid Feature Selection (HFS)

DR Prediction Model	**Accuracy**	**Precision**	**F1 Score**	**Recall**
XGboost	0.698	0.471	0.400	0.348
Random Forests	0.701	0.617	0.358	0.252
Elastic Net	0.710	0.667	0.206	0.122
SVM	0.678	0.800	0.075	0.040
KNN	0.662	0.423	0.323	0.261
XGboost+HFS	**0.729**	0.623	**0.457**	**0.361**
Random Forests+HFS	0.727	0.705	0.380	0.261
Elastic Net+HFS	0.718	0.769	0.276	0.168
SVM+HFS	0.694	**0.857**	0.095	0.050
KNN+HFS	0.646	0.403	0.290	0.227

Tables 5 and 6 compare the performance of several complication prediction models using the feature set selected by Hybrid Feature Selection (HFS) with the full set of 38 variables from the original dataset. These comparisons show that the clinical markers identified by Hybrid Feature Selection (HFS) are significantly associated with complications and can improve the performance of several prediction models.

Table 5 delineates the performance of diverse DKD prediction models pre- and post-application of Hybrid Feature Selection (HFS). Findings underscore that all models witnessed varying degrees of enhancement subsequent to HFS. Particularly noteworthy are the pronounced advancements observed in the XGBoost, Random Forest, and SVM models. Notable metrics improvements include XGBoost's accuracy ascending from 0.856 to 0.902, its F1 score from 0.647 to 0.704, and its recall from 0.531 to 0.613. The Random Forest model demonstrated a significant precision escalation from 0.712 to 0.894, while other metrics exhibited marginal alterations. SVM exhibited the most conspicuous overall improvement, with accuracy escalating from 0.823 to 0.883, precision from 0.735 to 0.810, F1 score from 0.554 to 0.667, and recall from 0.444 to 0.567. Collectively, XGBoost attained preeminence in accuracy, F1 score, and recall, while Random Forest excelled in precision.

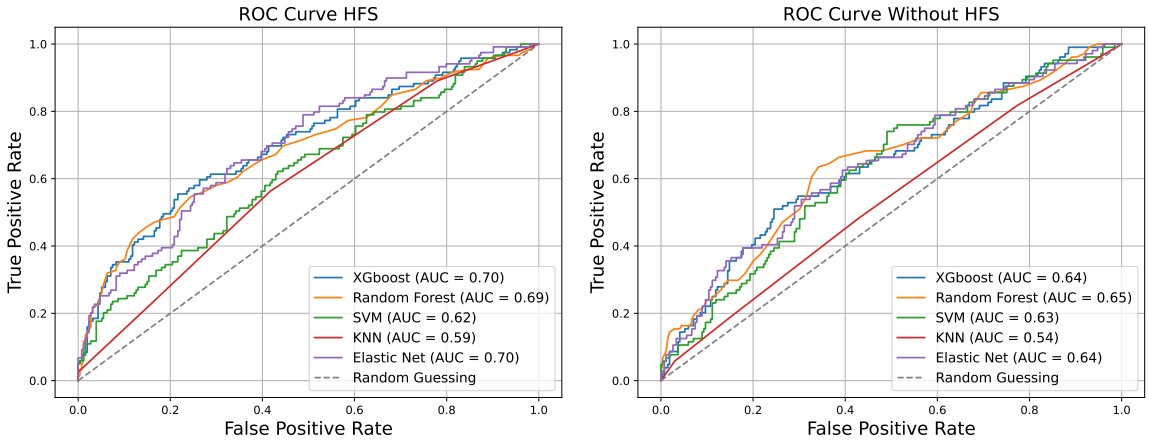
Table 6 shows the comparative performance of different DR prediction models before and after the application of HFS. While the overall predictive effectiveness of the DR models lags behind that of the DKD models, their performance improves to varying degrees after HFS integration. Specifically, the XGBoost model shows an improvement in accuracy from 0.698 to 0.729, precision from 0.471 to 0.623, F1 score from 0.400 to 0.457, and recall from 0.348 to 0.361. Similarly, the Random Forest model shows an increase in accuracy from 0.701 to 0.727 and precision from 0.617 to 0.705, while showing marginal differences in the other two metrics. It is worth noting the superior performance of XGBoost in terms of accuracy, F1 score, and recall. Although the SVM achieves the highest precision of 0.857, its F1 score and recall remain significantly low.



(a) Model with HFS Features Set

(b) Model without HFS Features Set

Figure 5: DKD Prediction Models ROC



(a) Model with HFS Features Set

(b) Model without HFS Features Set

Figure 6: DR Prediction Models ROC

Figures 5a and 5b show the ROC curves and corresponding AUC values for DKD prediction models after hybrid feature selection (HFS) and without HFS, respectively. AUC serves as a central metric for evaluating model effectiveness. Notably, there is a noticeable improvement in AUC for all models after HFS implementation. Specifically, among the best performing models, XGBoost shows an increase from 0.86 to 0.89, Random Forest from 0.83 to 0.88, and Elastic Net from 0.81 to 0.83.

Figures 6a and 6b show the ROC curves for the DR prediction models. The AUC values for the XGBoost, Random Forest, and Elastic Net models without HFS hover around 0.65. However, after implementing HFS, the AUC for each of these models increases to approximately 0.70.

Integrating the performance of the models from Tables 5 and 6, as well as Figures 5 and 6, we observe significant improvements in predictive performance after hybrid feature selection (HFS). Considering the comprehensive evaluation of model metrics and AUC, the XGBoost model emerges as the best choice for predicting both DKD and DR.

3.2 Causal Inference

3.2.1 Design Causal Questions

The final results of feature selection include several widely used biomarkers such as BMI, SBP, SCR, Insulin, and Sex (cite systematic). Additionally, there are some newly discovered markers in recent years, such as UA [14], CCB [1], ARB [1], and ALB [3]. However, the table also includes two less common indicators: C-peptide and TBIL. C-peptide is a crucial measure of pancreatic beta cell function due to its stable test window and association with diabetes type and disease duration.[15] Some articles have also recently proposed that C-peptide may have an impact on diabetic kidney disease. [25] [10]. However, current research on c-peptide faces limitations, such as challenges in interpreting results to predict clinical outcomes and the lack of widespread adoption in clinical guidelines despite its potential diagnostic and prognostic value. [17] [21] Total Bilirubin (TBIL) is also a rarely studied indicator for the complication DR. Some studies suggest that TBIL

has antioxidant properties, and low bilirubin concentrations may increase the risk of cardiovascular disease [16] and various other diseases [23]. However, these studies face many challenges. According to Vitek's research, clinical data on TBIL metabolism is often confounded by factors such as inappropriate reference ranges, lack of standardized TBIL measurement methods, and the absence of appropriately designed large epidemiological studies. Therefore, TBIL and C-peptide are included as study indicators.[23]

When estimating the causal effect of changes in the biomarkers C-peptide and TBIL on complications, it is important to first consider these changes as interventions. This requires setting a threshold for the biomarkers. Since the XGBoost model showed the best predictive performance during HFS, and since XGBoost is a decision tree-based model that assigns a split threshold for each feature, we extracted the split thresholds for C-peptide and TBIL. After consulting with physicians, we set the threshold for C-peptide at 1.35 ng/mL and for TBIL at 12.93 µmol/L.

Since C-peptide is negatively correlated with DKD risk, i.e. lower C-peptide levels correspond to higher DKD risk, we consider C-peptide levels below the threshold as "intervention" (A=1). Similarly, TBIL is negatively correlated with DR risk, so we consider TBIL levels below 12.93 µmol/L as "intervention" (A=1).

According to the structure shown in Figure 4, we have identified two backdoor paths that need to be closed. Since U is typically difficult to measure, we need to control for the two types of confounders, $L1$ and $L2$. Based on the literature review and consultation with medical professionals, we conducted research on the confounding between the TBIL $\rightarrow$ DR and the C-peptide $\rightarrow$ DKD.

To identify confounding factors in the relationship between TBIL $\rightarrow$ DR . Proper control of pre-analytical variables is crucial when assessing serum bilirubin concentrations, as stated in Vitek's paper[23]. Furthermore, many studies indicate an association between indicators related to blood pressure and TBIL. Additionally, Lin suggests that bilirubin is related to BMI, SBP, and other indices, which are also significantly associated with DR and could potentially produce confounding bias. [16]. Therefore, in the process of estimating TBIL $\rightarrow$ DR, we categorize indicators related to blood pressure, such as SBP and CCB, as $L1$. Some objective signs, such as BMI, AGE, Smoking, Insulin, alcohol, and sex, are categorized as $L2$. We simultaneously conditional on both $L1$ and $L2$.

Next, we identify the confounding effects between C-peptide $\rightarrow$ DKD. In the yaribeygi's study, the experiment controlled for BMI and other medication use as confounding factors.[25] Hills' research also mentioned the relationship between C-peptide and blood glucose. [10] So, GLUC, which represents blood glucose, was selected as $L1$ under the guidance of medical professionals. Other objective indicators, such as BMI, insulin, alcohol, and smoking, were selected as $L2$ and conditioned.

3.2.2 Risk Difference and Risk Ratio

The study estimated the impact of elevated C-peptide on the risk of diabetic kidney disease (DKD) and the impact of elevated total bilirubin (TBIL) on the risk of diabetic retinopathy (DR) using both augmented inverse probability weighting (AIPW) and flexible doubly robust estimation (FLAME) methods. Relevant confidence intervals were also obtained.

Table 7: Risk Difference measured by different methods

Marker	Complication	Methods	Risk Difference	Confidence Interval
TBIL	DR	FLAME	5.1%	(1.1%, 9.1%)
		AIPW	5.0%	(1.0%, 9.1%)
C-peptide	DKD	FLAME	13.3%	(9.6%,17.0%)
		AIPW	13.7%	(10%, 17.3%)

Table 8: Risk Ratio measured by different methods

Marker	Complication	Methods	Risk Ratio	Confidence Interval
TBIL	DR	FLAME	1.26	(1.05,1.53)
		AIPW	1.27	(1.07, 1.58)
C-peptide	DKD	FLAME	1.88	(1.57, 2.27)
		AIPW	1.91	(1.68, 2.34)

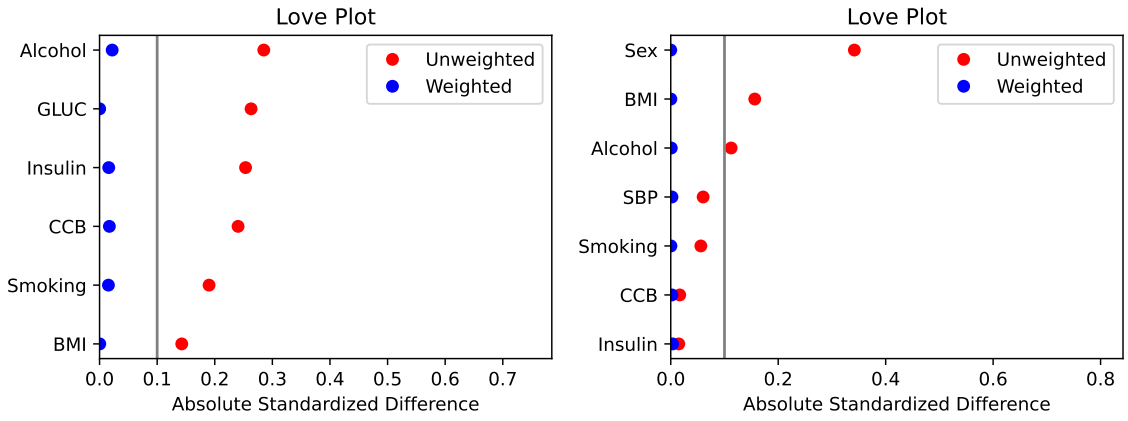
Two metrics were employed to evaluate the causal effect of biochemical indicators on complications: risk difference (RD) and risk ratio (RR). Table 7 illustrates that the estimated risk difference for C-peptide on DKD is over 13%, while for TBIL on DR, it is approximately 5%. This can be interpreted as follows: if all diabetic patients had C-peptide levels below 1.35 ng/ml, the probability of developing DKD would be expected to increase by 13 percentage points compared to those with C-peptide levels above 1.35 ng/ml. Conversely, if all diabetic patients had TBIL levels below 12.93 umol/L, the risk of developing DR would be expected to increase by 5 percentage points compared to those with TBIL levels above 12.93 umol/L. It is notable that the results estimated by both the AIPW and FLAME methods are highly consistent, which reinforces the reliability of our findings.

The results of the risk ratio (RR) presented in Table 8 can be interpreted as follows: if all diabetic patients had C-peptide levels below 1.35 ng/ml, the likelihood of developing diabetic kidney disease (DKD) would be approximately 1.9 times that of patients with C-peptide levels above 1.35 ng/ml. Conversely, if all diabetic patients had TBIL levels below 12.93 umol/L, the risk of developing DR would be approximately 1.26 times that of patients with TBIL levels above 12.93 umol/L.

3.2.3 Diagnosis of Causal Inference Results

Three types of diagnostic plots are used for AIPW inference:

The Love plot displays the differences before and after weighting for the intervention group (A=1) and the non-intervention group (A=0) variables. The blue points represent the differences after variable weighting. If all the blue points are on the left side of the gray line segment, it indicates that the weighting effect is good, and there is not much difference between the treatment and non-treatment groups in terms of covariates.

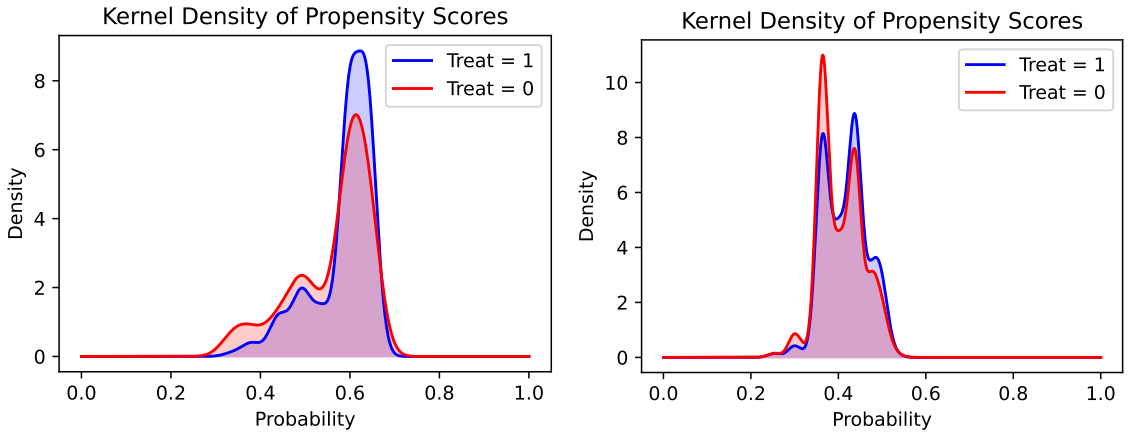


(a) C-peptide AIPW love plot

(b) TBIL AIPW love plot

Figure 7: The LOVE plot: Examining whether there are differences in covariates between the control group and the intervention group.

In the LOVE plot 7, the blue dots represent covariates after weighting, the red dots represent covariates before weighting, and the gray line represents a standardized difference of 0.1 between the intervention and control groups. If the blue dots after covariate weighting are located to the left of the gray vertical line, it indicates that there is minimal covariate difference between the intervention and control groups. In Figures 7a and 7b, it can be observed that all blue dots are to the left of the gray line, suggesting that after weighting, there is minimal covariate difference between the intervention and control groups in the inference of TBIL's effect on DR and C-peptide's effect on DKD. This implies that the results obtained are minimally influenced by covariates.

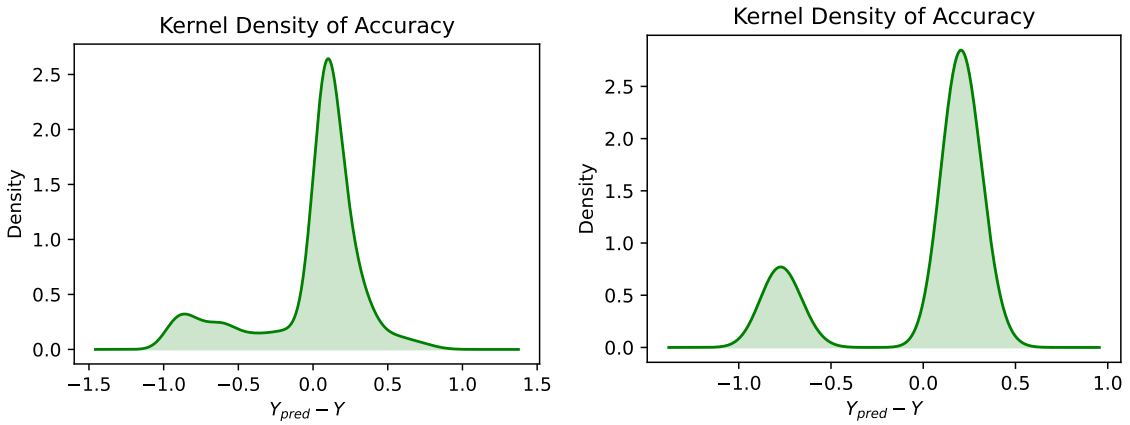


(a) C-peptide propensity score

(b) TBIL propensity score

Figure 8: Propensity Score Distribution plot: Describing the distribution of propensity scores to check the Positivity assumption.

Figure 8 is the propensity score distribution plot. This plot describes the distribution of propensity scores for the intervention and non-intervention groups. As Figures 8a and 8b demonstrate, the distributions of propensity scores for the intervention and control groups exhibit a high degree of overlap. This indicates that the positivity assumption is met during the causal inference process.



(a) C-peptide prediction distribution

(b) TBIL prediction distribution

Figure 9: The Prediction Distribution Plot: Check extreme outliers during the prediction process

Figure 9 depicts the distribution of residuals for the prediction of complication models in the AIPW process. The tails on both sides of the curves represent extreme outliers. Figures 9a and 9b show that both distribution curves have few variables in their tails, indicating that extreme outliers do not affect the causal inference process.

3.3 Discussion

This study contributes to the field of diabetic microvascular complications by identifying two novel non-traditional biomarkers: C-peptide and TBIL. C-peptide has been identified as a biomarker for DKD. By integrating machine learning and causal inference analysis, it can be concluded that if all patients' C-peptide levels are below 1.35 ng/ml, their risk of developing DKD is approximately 13% higher compared to when their C-peptide levels are above 1.35 ng/ml. Similarly, TBIL has been identified as a biomarker for DR. It can be reasonably assumed that if all patients' TBIL levels are below 12.93 umol/L, their risk of developing DR is expected to be approximately 5% higher compared to when their TBIL levels are above 12.93 umol/L.

Secondly, this study compares the differences between the biomarkers selected for different complications, thereby providing a reference for further pathological investigation. Among the 19 biomarkers identified as being related to DKD, in addition to traditional diabetes indicators such as insulin and serum creatinine (SCr), DKD was also significantly associated with certain blood pressure indicators such as ARB and SBP, as well as the uric acid (UA) indicator. In contrast, the 10 biomarkers identified for DR did not include clinical indicators significantly related to blood pressure and uric acid. This suggests that hypertension and hyperuricemia increase the risk of DKD from a pathological perspective, whereas their impact on DR is less significant. However, DR showed

a notable correlation with bilirubin indicators, including total bilirubin (TBIL), direct bilirubin (DBIL), and indirect bilirubin (IBIL). To enhance the ability to predict the occurrence of DR, further investigation into the pathological mechanisms of bilirubin is warranted. This study explores the potential pathological mechanisms by comparing the related variables of the two complications, providing new directions for future research and clinical applications.

With regard to methodology, this study also presents innovations. Traditional randomized controlled trials (RCTs) combined with statistical analysis require strict experimental conditions and often involve small sample sizes, making them more suitable for evaluating established biomarkers. In contrast, our approach, combining machine learning with causal inference, can be applied to large datasets and is suitable for discovering non-traditional biomarkers, providing guidance for future clinical trials. In the selection of biomarkers, we employed a hybrid feature selection model that integrated various statistical and machine learning methods. The selected biomarkers exhibited significant correlations with the respective complications and played a crucial role in predicting these complications. When assessing the impact of biomarkers on complications, given the absence of RCT conditions, we utilized a machine learning combined with causal inference model to control confounding factors, simulating RCT conditions. It is also noteworthy that the majority of the selected biomarkers have been previously mentioned in recent studies, thereby further validating the effectiveness and reliability of our approach.

References

[1] G Bakris and D White. Effects of an ace inhibitor combined with a calcium channel blocker on progression of diabetic nephropathy. *Journal of human hypertension*, 11(1):35–38, 1997.

[2] Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. Inference on treatment effects after selection among high-dimensional controls. *Review of Economic Studies*, 81(2):608–650, 2014.

[3] Satilmis Bilgin, Ozge Kurtkulagi, Burcin Meryem Atak Tel, Tuba Taslamacioglu Duman, Gizem Kahveci, Atiqa Khalid, and Gulali Aktas. Does c-reactive protein to serum albumin ratio correlate with diabetic nephropathy in patients with type 2 diabetes mellitus? the care time study. *Primary care diabetes*, 15(6):1071–1074, 2021.

[4] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

[5] James Dean Brown. Point-biserial correlation coefficients. *Statistics*, 5(3):12–6, 2001.

[6] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[7] Joakim Ekström. The phi-coefficient, the tetrachoric correlation coefficient, and the pearson-yule debate. 2011.

[8] Laura Freijeiro-González, Manuel Febrero-Bande, and Wenceslao González-Manteiga. A critical review of lasso and its derivatives for variable selection under dependence among covariates. *International Statistical Review*, 90(1):118–145, 2022.

[9] Pablo M Granitto, Cesare Furlanello, Franco Biasioli, and Flavia Gasperi. Recursive feature elimination with random forest for ptr-ms analysis of agroindustrial products. *Chemometrics and intelligent laboratory systems*, 83(2):83–90, 2006.

[10] Claire E Hills, Nigel J Brunskill, and Paul E Squires. C-peptide as a therapeutic tool in diabetic nephropathy. *American Journal of Nephrology*, 31(5):389–397, 2010.

[11] Jin Huang and Charles X Ling. Using auc and accuracy in evaluating learning algorithms. *IEEE Transactions on knowledge and Data Engineering*, 17(3):299–310, 2005.

[12] Stefano M Iacus, Gary King, and Giuseppe Porro. Multivariate matching methods that are monotonic imbalance bounding. *Journal of the American Statistical Association*, 106(493):345–361, 2011.

[13] Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge university press, 2015.

[14] Diana I Jalal, David M Maahs, Peter Hovind, and Takahiko Nakagawa. Uric acid as a mediator of diabetic nephropathy. In *Seminars in nephrology*, volume 31, pages 459–465. Elsevier, 2011.

[15] Emma Leighton, Christopher AR Sainsbury, and Gregory C Jones. A practical review of c-peptide testing in diabetes. *Diabetes therapy*, 8:475–487, 2017.

[16] Jing-Ping Lin, Libor Vitek, and Harvey A Schwertner. Serum bilirubin and genes controlling bilirubin concentrations as biomarkers for cardiovascular disease. *Clinical chemistry*, 56(10):1535–1543, 2010.

[17] NICE. Diabetes in adults. list of quality statements. guidance and guidelines, 2016. `https://www.nice.org.uk/guidance/qs6`, Last accessed on 2016.

[18] Paul R Rosenbaum, P Rosenbaum, and Briskman. *Design of observational studies*, volume 10. Springer, 2010.

[19] Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

[20] Cynthia Rudin, Chaofan Chen, Zhi Chen, Haiyang Huang, Lesia Semenova, and Chudi Zhong. Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistic Surveys*, 16:1–85, 2022.

[21] David B Sacks, Mark Arnold, George L Bakris, David E Bruns, Andrea Rita Horvath, M Sue Kirkman, Ake Lernmark, Boyd E Metzger, and David M Nathan. Position statement executive summary: guidelines and recommendations for laboratory analysis in the diagnosis and management of diabetes mellitus. *Diabetes care*, 34(6):1419–1423, 2011.

[22] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288, 1996.

[23] Libor Vitek. Bilirubin as a predictor of diseases of civilization. is it time to establish decision limits for serum bilirubin concentrations? *Archives of biochemistry and biophysics*, 672:108062, 2019.

[24] Tianyu Wang, Marco Morucci, M Usaid Awan, Yameng Liu, Sudeepa Roy, Cynthia Rudin, and Alexander Volfovsky. Flame: A fast large-scale almost matching exactly approach to causal inference. *Journal of Machine Learning Research*, 22(31):1–41, 2021.

[25] Habib Yaribeygi, Mina Maleki, Thozhukat Sathyapalan, and Amirhossein Sahebkar. The effect of c-peptide on diabetic nephropathy: A review of molecular mechanisms. *Life sciences*, 237:116950, 2019.