Black Sea Journal of Engineering and Science

doi: 10.34248/bsengineering.1880946

Open Access Journal
e-ISSN: 2619 – 8991

Research Article
Volume 9 - Issue 3: 1301-1311 / May 2026

INFLUENCE OF CONFOUNDING CONTROL ON LIVER FIBROSIS MODELING: A PROPENSITY SCORE MATCHING AND LASSO-BASED LOGISTIC REGRESSION APPROACH

Fulden CANTAŞ TÜRKİŞ[1]*, Buğra VAROL[1]

[1]*Muğla Sıtkı Koçman University, Faculty of Medicine, Department of Biostatistics, 48000, Muğla, Türkiye*

Abstract: In large-scale observational datasets, associations between clinical outcomes and biomarkers are highly sensitive to confounding control and statistical modeling strategies. This study aimed to evaluate how confounding-control and variable-selection strategies influence the observed association between retinopathy and liver fibrosis in a large observational dataset, thereby providing a methodological demonstration of how statistical modeling decisions may influence epidemiological inference. Data were obtained from the National Health and Nutrition Examination Survey (NHANES) 2005-2008 cycles. Liver fibrosis was defined using the FIB-4 index and classified as no/mild fibrosis (FIB-4 < 1.45) or significant fibrosis/cirrhosis (FIB-4 ≥ 1.45). Retinopathy status and severity were determined using standardized retinal imaging data. To improve baseline comparability between fibrosis groups, propensity score matching (PSM) was implemented based on age and sex, and covariate balance was assessed using standardized mean differences. Variable selection was performed using the Least Absolute Shrinkage and Selection Operator (LASSO) with 10-fold cross-validation, and the optimal penalty parameter (λ) was selected based on the minimum cross-validated error (λ_{min}), followed by multivariable logistic regression modeling. Variables used in the calculation of the FIB-4 index were excluded from regression analyses to avoid circular inference. A total of 5,364 participants were included. Before matching, substantial imbalance was observed between fibrosis groups, particularly for age and sex. After 1:1 propensity score matching, adequate covariate balance was achieved. LASSO-based variable selection identified hepatitis C virus infection, body mass index, height, race, and retinopathy as candidate predictors of liver fibrosis. In the final multivariable logistic regression model, hepatitis C virus infection showed the strongest association with significant liver fibrosis (OR = 2.70, 95% CI: 1.66-4.39), while body mass index and height demonstrated modest but statistically significant associations. Retinopathy was not independently associated with liver fibrosis after multivariable adjustment. The results demonstrate that the apparent association between retinopathy and liver fibrosis in observational data is highly dependent on statistical modeling choices and confounding-control strategies. Rather than supporting a direct clinical relationship, the findings emphasize how analytical design and variable selection methods can substantially shape conclusions derived from large-scale health datasets. From a methodological perspective, this study illustrates how confounding-control and variable-selection strategies can alter epidemiological inference in observational research. This study highlights the importance of transparent and rigorously justified statistical modeling frameworks in applied data-driven research.

Keywords: Propensity score matching, LASSO regression, Logistic regression, Confounding, Covariate balance

***Corresponding author:** Muğla Sıtkı Koçman University, Faculty of Medicine, Department of Biostatistics, 48000, Muğla, Türkiye
E mail: fuldencantas@mu.edu.tr (F. CANTAŞ TÜRKİŞ)
Fulden CANTAŞ TÜRKİŞ https://orcid.org/0000-0002-7018-7187
Buğra VAROL https://orcid.org/0000-0001-8052-7782

Received: February 03, 2026
Accepted: April 25, 2026
Published: May 15, 2026

Cite as: Cantaş Türkiş, F., & Varol, B. (2026). Influence of confounding control on liver fibrosis modeling: a propensity score matching and lasso-based logistic regression approach. *Black Sea Journal of Engineering and Science, 9*(3), 1301-1311.

1. Introduction

In observational epidemiological studies, the evaluation of relationships between clinical outcomes and biological markers critically depends on appropriate control of confounding factors and the selection of suitable statistical modeling strategies. Particularly in studies conducted within the context of metabolic disorders, the coexistence of multiple risk factors and the complex interrelationships among these factors render conventional regression approaches increasingly limited. These challenges become more pronounced in the investigation of chronic and multifactorial outcomes such as liver fibrosis, where methodological rigor is essential to ensure the validity and interpretability of findings.

Methodological complexities become particularly evident when composite clinical indices are used to characterize disease severity. The Fibrosis-4 (FIB-4) index is a widely used non-invasive tool for assessing liver fibrosis and is calculated using age, aspartate aminotransferase (AST), alanine aminotransferase (ALT), and platelet count. Due to its simplicity and accessibility, FIB-4 has been extensively applied in large-scale epidemiological datasets (Chhabra et al., 2022; Shaji et al., 2022; Woodard and Abrams, 2024). Nevertheless, despite these advantages, the use of composite indices in analytical modeling may introduce important statistical challenges,

particularly when component variables are reintroduced into regression analyses. This practice may result in circular inference and biased effect estimates, thereby complicating model interpretation and potentially obscuring the true relationship between clinical variables and disease outcomes.

Within this context, several studies have examined the relationship between metabolic dysfunction–associated liver disease (MASLD) and diabetic microvascular complications, particularly retinopathy. Most of these investigations have relied on conventional multivariable regression frameworks (Asero et al., 2023; Jacob et al., 2023; Erman et al., 2024; Li et al., 2024; Mantovani et al., 2024). For example, Jacob et al. (2023) evaluated this association in a relatively small clinical cohort, while Mantovani et al. (2024) analyzed MASLD and fibrosis categories among adults with diabetes using predefined logistic regression models. Similarly, Li et al. (2024) used NHANES data to assess whether diabetic retinopathy could predict significant hepatic fibrosis using correlation, logistic regression, and ROC analyses. However, when exposure and outcome variables share multiple common covariates, conventional regression adjustment may not fully eliminate residual confounding. Consistent with this concern, prior studies have reported heterogeneous findings, with some identifying significant associations while others observed attenuation after multivariable adjustment (Jacob et al., 2023; Li et al., 2024; Mantovani et al., 2024). These inconsistencies suggest that an important methodological question remains unresolved: whether the retinopathy–fibrosis association remains robust when alternative confounding-control and variable-selection strategies are applied within the same observational dataset.

To address confounding more effectively than conventional multivariable adjustment, propensity score-based methods have emerged as an alternative strategy in observational research. Propensity score matching (PSM) aims to improve covariate balance between comparison groups by matching individuals with similar probabilities of exposure, thereby enhancing comparability (Jeong and Kim, 2024; Burgos-Ochoa and Clouth, 2025). By reducing systematic differences between groups before modeling, PSM allows for more reliable evaluation of associations and approximates the balance typically achieved in randomized study designs.

Variable selection constitutes another key component of statistical modeling in observational studies. In settings with a large number of potential covariates, traditional stepwise regression approaches are prone to model instability and overfitting. The Least Absolute Shrinkage and Selection Operator (LASSO) regression method addresses these limitations by applying L1 regularization, shrinking regression coefficients and eliminating weak predictors to yield more parsimonious and generalizable models (Schonlau, 2023; Bangchang, 2024). Previous studies have demonstrated that LASSO-based approaches may offer superior discriminative performance compared with traditional fibrosis indices in predicting liver fibrosis (Feng et al., 2021; Guo et al., 2023; Zhang et al., 2023). In addition to improving predictive performance, LASSO enables objective variable selection in complex epidemiological datasets and reduces the risk of overfitting when multiple potential confounders are considered simultaneously.

NHANES is a large, population-based epidemiological dataset that provides comprehensive demographic, clinical, and laboratory information collected using standardized protocols across survey cycles (Ellis and Souza, 2021; Bo and Yang, 2025). Its large sample size and rich variable structure allow for the application of advanced confounding-control and variable-selection strategies in observational research (Cheang et al., 2022; Issanov et al., 2022; Storz, 2023).

Against this background, few studies have systematically examined how different confounding-control and variable-selection strategies influence the observed association between retinopathy and liver fibrosis within the same epidemiological dataset. To our knowledge, no previous study has simultaneously applied PSM and LASSO-based variable selection within NHANES data to examine the robustness of the retinopathy–liver fibrosis association under alternative confounding-control frameworks. In the present study, NHANES 2005-2008 data were used to examine how the association between retinopathy and liver fibrosis, assessed using the FIB-4 index, changes after applying PSM to improve covariate balance and LASSO-based regularization to guide variable selection. By integrating these two complementary analytical strategies within a single epidemiological framework, this study aims to provide a methodological demonstration of how confounding-control and variable-selection approaches can influence epidemiological interpretations derived from large observational health datasets.

2. Materials and Methods

2.1. Data Source

This study was conducted using data from the National Health and Nutrition Examination Survey (NHANES), a publicly available database maintained by the National Center for Health Statistics (NCHS) of the Centers for Disease Control and Prevention (CDC). Data from the 2005-2008 survey cycles were included in the analysis. NHANES is a large-scale, cross-sectional survey designed to provide nationally representative information on the health and nutritional status of the civilian, non-institutionalized population of the United States. All participants provided written informed consent at the time of data collection.

The present study represents a secondary analysis of fully anonymized public data; therefore, additional ethical approval was not required.

2.2. Study Population

Participants were included if they had available retinal imaging data as well as complete demographic,

anthropometric, and laboratory information required for the assessment of liver fibrosis. Individuals with missing data in key variables were excluded using a complete-case approach. After applying the inclusion and exclusion criteria, a total of 5364 participants were eligible and included in the final study population.

2.3. Assessment of Retinopathy

Retinopathy status was determined based on retinal imaging data collected according to the standardized NHANES protocol. Retinopathy severity was classified into four categories: no retinopathy, mild non-proliferative diabetic retinopathy (NPDR), moderate/severe NPDR, and proliferative retinopathy. For regression analyses, retinopathy was additionally evaluated as a binary variable (presence vs. absence).

2.4. Definition of Liver Fibrosis

Liver fibrosis was assessed using the Fibrosis-4 (FIB-4) index, a widely used non-invasive marker calculated based on age, aspartate aminotransferase (AST), alanine aminotransferase (ALT), and platelet count. Participants were classified as having no or mild fibrosis (FIB-4 < 1.45) or significant liver fibrosis/cirrhosis (FIB-4 ≥ 1.45). Variables included in the calculation of the FIB-4 index (AST, ALT, and platelet count) were presented for descriptive purposes only and were excluded from regression analyses to avoid circularity.

2.5. Propensity Score Matching

As this was an observational study, baseline differences between fibrosis groups were expected. To reduce potential confounding and improve comparability between groups, PSM (Rosenbaum and Rubin, 1983; Austin, 2011) was applied. The propensity score was defined as the conditional probability of having significant liver fibrosis given observed covariates. Propensity scores were estimated using a logistic regression model including age and sex, which were selected a priori as clinically relevant confounders and showed substantial imbalance between groups prior to matching.

Matching was performed using the nearest-neighbor method with a 1:1 matching ratio and without replacement. A caliper width of 0.15 standard deviations of the logit of the propensity score was applied to prevent poor-quality matches. Individuals who could not be adequately matched within the specified caliper were excluded from the matched cohort. Covariate balance before and after matching was evaluated using standardized mean differences (SMDs), with values below 0.10 considered indicative of adequate balance. Accordingly, PSM was employed as a methodological tool to enhance group comparability and support subsequent model-based analyses, rather than to establish causal relationships.

2.6. Statistical Analysis

Continuous variables were summarized as median (minimum-maximum) due to non-normal distributions, while categorical variables were presented as frequencies and percentages. The normality of continuous variables was assessed using the Kolmogorov-Smirnov test. Comparisons between groups were performed using the Mann-Whitney U test for continuous variables.

To identify independent factors associated with significant liver fibrosis, variable selection was performed using the Least Absolute Shrinkage and Selection Operator (LASSO) method. The optimal penalty parameter was determined using 10-fold cross-validation. Variables selected by the LASSO procedure were subsequently entered into a multivariable logistic regression model. Multicollinearity among predictors was assessed using variance inflation factors (VIFs). In regression analyses conducted after PSM, robust standard error estimates clustered at the matching subclass level were used to account for the dependence structure induced by the matching procedure.

To explore potential non-linear associations between body mass index (BMI) and liver fibrosis, restricted cubic spline regression was performed within the multivariable logistic regression framework.

PSM was implemented using the MatchIt package, and covariate balance was assessed using the cobalt package; graphical visualizations were generated using ggplot2 and gridExtra. LASSO regression was performed using the glmnet package, while additional data handling and diagnostics were conducted using dplyr, car, and haven packages.

All statistical analyses were performed using R software (version 4.5.1; R Foundation for Statistical Computing, Vienna, Austria). A two-sided p-value < 0.05 was considered statistically significant.

3. Results

3.1. Baseline Characteristics of the Study Population before Propensity Score Matching

Baseline characteristics and standardized mean differences of the study population before PSM are presented in Table 1. A total of 3248 participants had no or mild hepatic fibrosis, while 2116 participants were classified as having significant liver fibrosis or cirrhosis.

Before matching, several baseline characteristics exhibited meaningful imbalance between groups, as assessed by standardized mean differences. Participants with significant liver fibrosis tended to be older and more frequently male, and differences were also observed in race distribution, anthropometric measures, metabolic status, and selected clinical characteristics.

In particular, demographic and anthropometric variables, as well as platelet count and diabetes status, showed notable baseline imbalance, whereas height and viral hepatitis status demonstrated minimal imbalance. Retinopathy prevalence was higher in the fibrosis group; however, imbalances across retinopathy severity categories were generally small.

3.2. Covariate Balance after Propensity Score Matching

Following 1:1 nearest-neighbor PSM with a caliper of

0.15 standard deviations of the logit of the propensity score, a matched cohort with balanced baseline characteristics was obtained. Covariate balance before and after matching is summarized in Table 2 and illustrated in Figure 1.

Before matching, substantial imbalance was observed between fibrosis groups with respect to age (SMD = 1.50) and sex (SMD = 0.25). The large baseline imbalance observed for age reflects the structural role of age (SMD=1.50) in the FIB-4 index calculation. After matching, SMD values for the covariates included in the matching model were reduced below the predefined threshold of 0.10, indicating adequate balance for age and sex.. The distribution of propensity scores before and after matching is shown in Figure 2, demonstrating improved overlap between groups after matching.

Table 1. Baseline characteristics of the study population before propensity score matching

Variables	No/mild hepatic fibrosis (n=3248)	Significant liver fibrosis/cirrhosis (n=2116)	SMD
Baseline information			
Age (years)	52 (40-85)	70 (40-85)	1.499
Sex [n (%)]			
Male	1470 (45.3)	1223 (57.8)	0.252
Female	1778 (54.7)	893 (42.2)	
Race [n (%)]			
Non-Hispanic white	602 (18.5)	251 (11.9)	0.183
Non-Hispanic black	267 (8.2)	109 (5.2)	0.119
Mexican American	1607 (49.5)	1300 (61.4)	0.241
Other Hispanic	647 (19.9)	407 (19.2)	0.018
Other race	125 (3.8)	49 (2.3)	0.088
Physical examination			
Weight (kg)	81.05 (36.30-371)	78.10 (35.90-177.40)	0.192
Height (cm)	166.70 (140.30-203.80)	167.90 (138.60-197.10)	0.050
BMI (kg/m2)	28.90 (15.69-130.21)	27.63 (13.36-57.31)	0.255
Clinical information			
PLT (10^9/L)	39 (8-702)	33 (4-819)	0.337
ALT (U/L)	22 (8-178)	24 (7-400)	0.241
AST (U/L)	219 (10-799)	175 (14-412)	0.220
Retinopathy severity [n (%)]			
No retinopathy (Level 1)	2893 (89.1)	1806 (85.3)	0.114
Mild NPDR (Level 2)	282 (8.7)	261 (12.3)	0.117
Moderate/severe NPDR (Level 3)	60 (1.8)	36 (1.7)	0.008
PR (Level 4)	13 (0.4)	13 (0.6)	0.029
Retinopathy [n (%)]			
No	2893 (89.1)	1806 (85.3)	0.114
Yes	355 (10.9)	310 (14.7)	
HCV infection [n (%)]			
No	3170 (97.6)	2051 (96.9)	0.043
Yes	78 (2.4)	65 (3.1)	
HBV infection [n (%)]			
No	3017 (92.9)	1952 (92.2)	0.026
Yes	231 (7.1)	164 (7.8)	
T2DM [n (%)]			
No	2659 (81.9)	1606 (75.9)	0.147
Yes	589 (18.1)	510 (24.1)	

BMI= body mass index, PLT= platelet count, ALT= alanine aminotransferase, AST= aspartate aminotransferase, NPDR= non-proliferative diabetic retinopathy, PR= proliferative retinopathy, HCV= hepatitis C virus, HBV= hepatitis B virus, T2DM= type 2 diabetes mellitus, SMD= standardized mean difference. Descriptive statistics are shown as n (%) or median (minimum-maximum). Variables used in the calculation of the fibrosis index (AST, ALT, and platelet count) are presented for descriptive purposes only and were not included in subsequent regression analyses to avoid circularity. SMD > 0.10 indicates meaningful imbalance.

Table 2. Covariate balance before and after propensity score matching

Variable	Before matching SMD	After matching SMD
Age	1.499	0.072
Sex	0.254	0.018

Standardized mean difference (SMD) < 0.10 indicates adequate covariate balance.

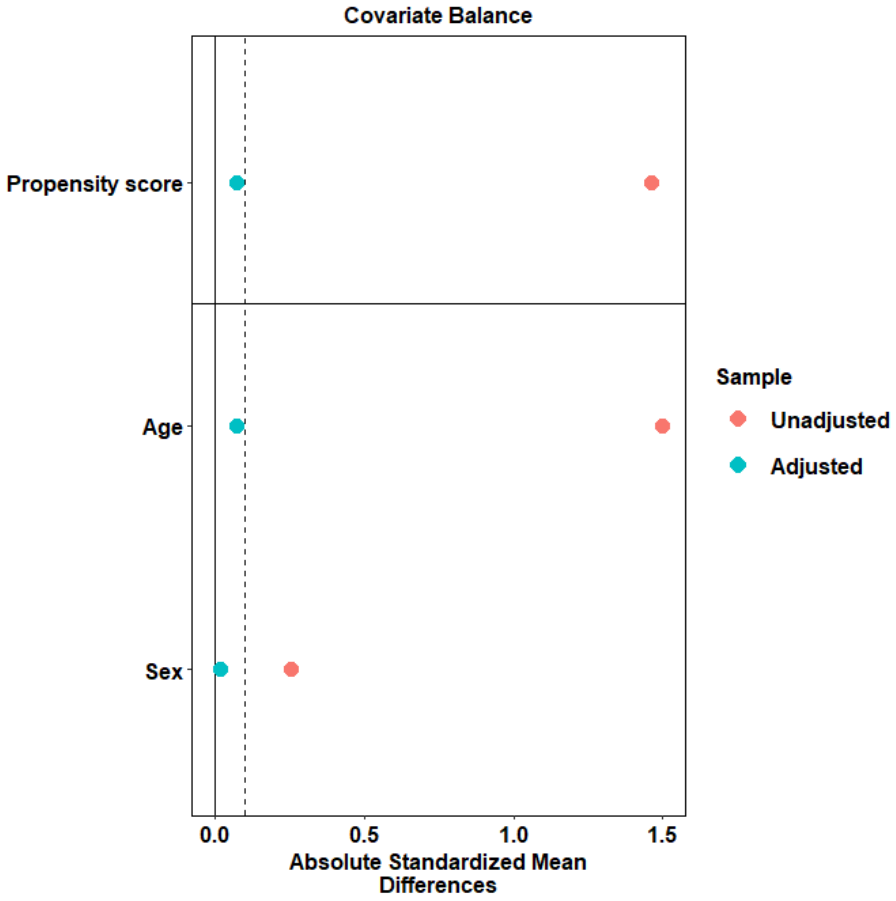

Figure 1. Covariate balance before and after propensity score matching. Absolute standardized mean differences (SMDs) for covariates before and after propensity score matching. The dashed line represents the balance threshold (SMD = 0.10).

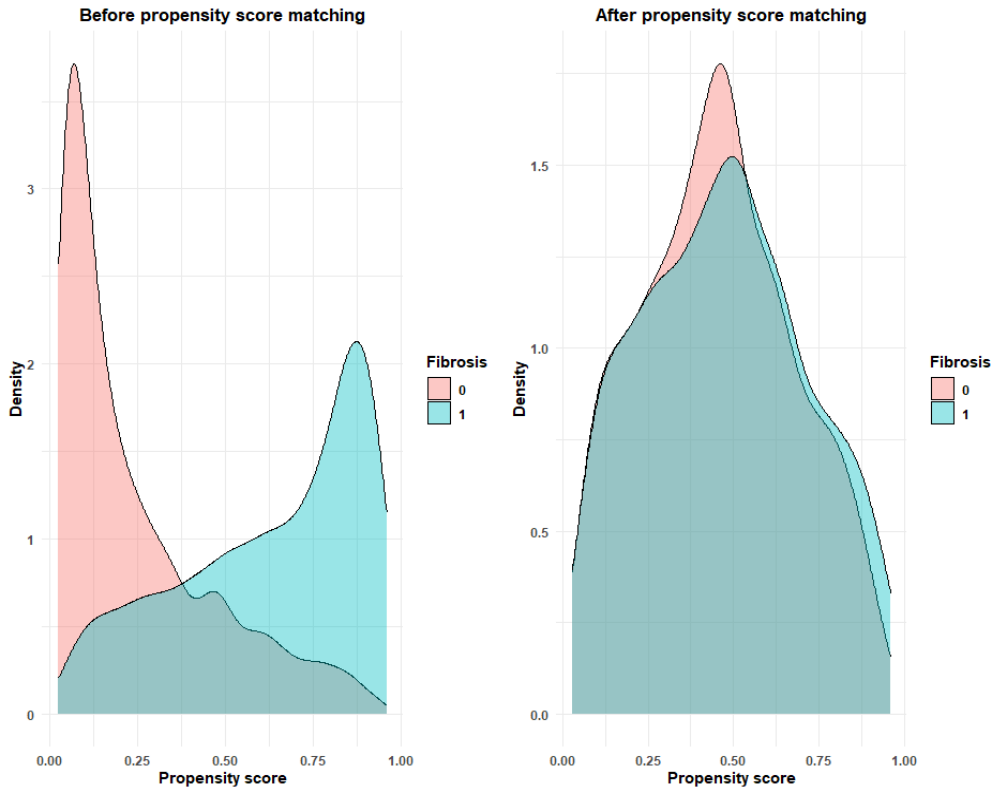

Figure 2. Distribution of propensity scores before and after propensity score matching. Fibrosis status was coded as 0 = no/mild hepatic fibrosis and 1 = significant liver fibrosis or cirrhosis.

3.3. Baseline Characteristics after Propensity Score Matching

Baseline characteristics and standardized mean differences of the matched cohort are presented in Table 3. After PSM, substantial improvement in baseline balance was achieved between the fibrosis groups.

Most demographic and clinical variables demonstrated adequate balance after matching, with standardized mean differences below the conventional threshold of 0.10, including age, sex, weight, retinopathy prevalence, retinopathy severity categories, and type 2 diabetes mellitus. Race distribution also showed acceptable balance across categories, with SMD values generally indicating small residual differences.

Anthropometric measures showed minimal residual imbalance, with height exhibiting SMD values close to the balance threshold, while body mass index demonstrated a modest imbalance. HBV was well balanced between groups, whereas HCV infection showed a small residual imbalance after matching.

Laboratory parameters used in the calculation of the fibrosis index, including ALT, AST, and platelet count, exhibited residual imbalance after matching and were therefore reported for descriptive purposes only and not included in subsequent regression analyses to avoid circularity.

3.4. Feature Selection Using LASSO Regression

To identify independent predictors of significant liver fibrosis while minimizing overfitting and addressing potential multicollinearity, variable selection was performed using the Least Absolute Shrinkage and Selection Operator (LASSO) approach. The optimal penalty parameter was determined as λ_{min} = 0.0057 through 10-fold cross-validation, corresponding to the minimum binomial deviance.

During the shrinkage process, Type 2 diabetes mellitus (T2DM) and body weight were penalized to zero, indicating a limited independent contribution in the presence of other covariates. The LASSO procedure retained five variables-HCV infection, body mass index (BMI), height, race, and retinopathy-as the most stable variables associated with liver fibrosis. The coefficient trajectories and cross-validation results are shown in Figure 3.

To assess the robustness of the selected variables, multicollinearity was evaluated using the Variance Inflation Factor (VIF). All VIF values were below 1.10 (Retinopathy: 1.01, BMI: 1.01, Race: 1.07, HCV: 1.00, and Height: 1.07), indicating negligible collinearity and supporting the suitability of these variables for inclusion in the final multivariable logistic regression model.

3.5. Factors Associated with Liver Fibrosis in the Matched Cohort

The results of the multivariable logistic regression analysis are presented in Table 4. The final multivariable logistic regression model, adjusted for variables selected by the LASSO procedure, demonstrated that HCV infection showed the strongest relationship with significant liver fibrosis/cirrhosis, with an odds ratio (OR) of 2.70 (95% CI: 1.66-4.39, P<0.001). Increasing

height was linked to a higher likelihood of liver fibrosis (OR: 1.01, 95% CI: 1.00-1.02, P=0.039), whereas body mass index (BMI) exhibited a modest but statistically significant inverse relationship with fibrosis status (OR: 0.98, 95% CI: 0.97-0.99, P=0.001).

Given the inverse association observed for BMI, restricted cubic spline regression was performed to evaluate potential non-linear associations. The spline model was adjusted for HCV infection, height, retinopathy, and race. The overall association between BMI and fibrosis remained statistically significant (P=0.007), whereas the test for non-linearity was not significant (P=0.447), suggesting an approximately linear relationship across the BMI range. In addition, BMI remained significantly different between fibrosis groups after propensity score matching (P=0.002), which is expected as BMI was not included in the propensity score model.

Although retinopathy was retained during the LASSO-based variable selection process, it did not demonstrate an independent relationship with liver fibrosis in the multivariable analysis (OR: 1.08, 95% CI: 0.86-1.36, P=0.494). Similarly, race did not show a statistically significant relationship with liver fibrosis after multivariable adjustment (P=0.0.097).

Table 3. Baseline characteristics of the study population after propensity score matching

Variables	No/mild hepatic fibrosis (n=1213)	Significant liver fibrosis/cirrhosis (n=1213)	SMD
Baseline information			
Age (years)	63 (40-85)	63 (40-85)	0.072
Sex [n (%)]			
Male	627 (51.7)	638 (52.6)	0.018
Female	586 (48.3)	575 (47.4)	
Race [n (%)]			
Non-Hispanic white	205 (16.9)	173 (14.3)	0.072
Non-Hispanic black	98 (8.1)	72 (5.9)	0.086
Mexican American	621 (51.2)	651 (53.7)	0.050
Other Hispanic	249 (20.5)	287 (23.7)	0.077
Other race	40 (3.3)	30 (2.5)	0.048
Physical examination			
Weight (kg)	80.30 (36.30-186.90)	79.50 (40.10-177.40)	0.061
Height (cm)	166.80 (141.50-195)	167.90 (138.60-197.10)	0.099
BMI (kg/m2)	28.94 (15.98-63.92)	28.16 (14.20-57.31)	0.130
Clinical information			
PLT (10^9/L)	33 (8-702)	41 (4-496)	0.173
ALT (U/L)	22 (10-104)	25 (7-329)	0.392
AST (U/L)	222 (10-556)	176 (14-400)	0.247
Retinopathy severity [n (%)]			
No retinopathy (Level 1)	1041 (85.8)	1038 (85.6)	0.006
Mild NPDR (Level 2)	131 (10.8)	141 (11.6)	0.025
Moderate/severe NPDR (Level 3)	35 (2.9)	25 (2.1)	0.052
PR (Level 4)	6 (0.5)	9 (0.7)	0.026
Retinopathy [n (%)]			
No	1041 (85.8)	1038 (85.6)	0.006
Yes	172 (14.2)	175 (14.4)	
HCV infection [n (%)]			
No	1191 (98.2)	1152 (95)	0.170
Yes	22 (1.8)	61 (5)	
HBV infection [n (%)]			
No	1116 (92)	1108 (91.3)	0.025
Yes	97 (8)	105 (8.7)	
T2DM [n (%)]			
No	905 (74.6)	922 (76)	0.033
Yes	308 (25.4)	291 (24)	

BMI= body mass index, PLT= platelet count, ALT= alanine aminotransferase, AST= aspartate aminotransferase, NPDR= non-proliferative diabetic retinopathy, PR= proliferative retinopathy, HCV= hepatitis C virus, HBV= hepatitis B virus, T2DM= type 2 diabetes mellitus, SMD= standardized mean difference. Variables used in the calculation of the fibrosis index (AST, ALT, and platelet count) are presented for descriptive purposes only and were not included in subsequent regression analyses to avoid circularity.

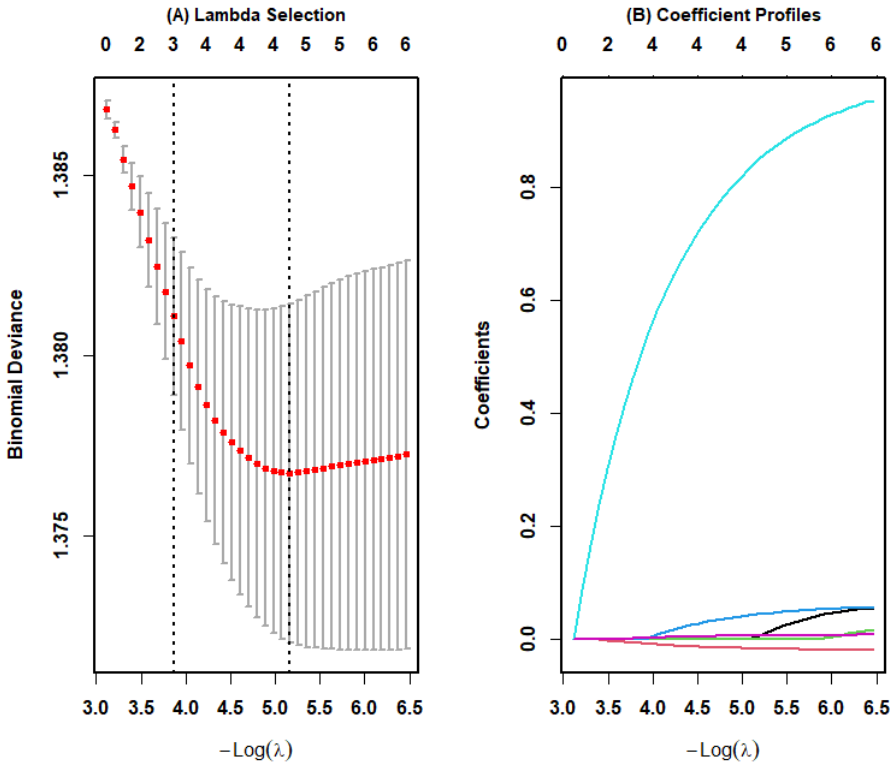

Figure 3. Cross-validated selection of the optimal penalty parameter (λ) and coefficient profiles obtained from the LASSO model. (A) The optimal penalty parameter log(λ) selection. The binomial deviance was plotted against log(λ). The left vertical dashed line represents the minimum error (λ_{min} = 0.0057), which was used to select the optimal features for the final model.

Table 4. Multivariable logistic regression results for liver fibrosis

Variables	β	OR	95% CI	p	VIF
HCV Infection (Yes)	0.995	2.70	(1.66 - 4.39)	<0.001	1.000
Height (cm)	0.007	1.01	(1.00 - 1.02)	0.039	1.070
BMI (kg/m2)	-0.021	0.98	(0.97 - 0.99)	0.001	1.010
Retinopathy (Yes)	0.080	1.08	(0.86 - 1.36)	0.494	1.010
Race (Ref: Non-Hispanic white)				0.097	
Non-Hispanic black	-0.130	0.878	(0.608 - 1.268)	0.488	
Mexican American	0.170	1.185	(0.932 - 1.506)	0.165	1.070
Other Hispanic	0.249	1.282	(0.974 - 1.688)	0.076	
Other race	-0.193	0.825	(0.490 - 1.386)	0.467	

Ref= reference category, β= regression coefficient, OR= odds ratio, CI= confidence interval; VIF= variance inflation factor. Model performance= Residual Deviance = 3326; AIC = 3338. Variables were selected using a LASSO regression-based procedure with an optimal penalty parameter of λ= 0.0057. Variables used in the calculation of the fibrosis index (AST, ALT, and platelet count) were excluded to avoid circularity.

4. Discussion

In this study, NHANES 2005-2008 data were used to examine the association between retinopathy and liver fibrosis, assessed using the FIB-4 index, within the framework of PSM and LASSO-based variable selection approaches. Rather than addressing clinical causality, the primary objective of this study was methodological: to examine how confounding-control strategies combined with regularized variable selection approaches influence epidemiological inference in large observational datasets. One of the main findings of this study is that the retinopathy-fibrosis association, which appeared statistically significant in unadjusted or minimally adjusted analyses, lost its statistical significance after comprehensive confounding control. This observation suggests that a substantial proportion of the conflicting results reported in the literature may stem from differences in analytical strategies. Indeed, while retinopathy has been reported as a significant predictor of fibrosis in some studies using NHANES data (Li et al., 2024), other investigations conducted in similar populations have shown that associations observed in univariate analyses either attenuate or disappear after multivariable adjustment, or fail to demonstrate an independent relationship (Deravi et al., 2023; Jacob et al.,

2023). Such heterogeneity is closely related to the extent of confounding control and the statistical modeling approaches employed. Importantly, the present findings demonstrate that conclusions drawn from large observational datasets may change substantially depending on how confounding control and variable selection are implemented. This highlights the importance of carefully evaluating analytical strategies when interpreting epidemiological associations.

In observational data, conventional multivariable regression models may be insufficient for adequately controlling confounding. In the present study, PSM was applied to achieve covariate balance between fibrosis groups, particularly with respect to key variables such as age and sex, thereby improving comparability. The primary purpose of PSM is not to produce direct causal effect estimates, but rather to reduce systematic differences between groups and to facilitate more consistent and less biased subsequent modeling steps. This approach is widely used in datasets such as NHANES, which lack randomization and involve multiple potential confounding factors (Jeong and Kim, 2024; Burgos-Ochoa and Clouth, 2025). Importantly, the propensity score model was deliberately restricted to age and sex, which represented the most pronounced sources of baseline imbalance between groups. This design choice reflects a methodological preference to use PSM for addressing major demographic confounding, while allowing additional clinical and metabolic factors to be evaluated within the variable selection and multivariable regression framework.

An important methodological consideration in this analysis relates to the role of age in the definition of the FIB-4 index. Because age is directly incorporated into the FIB-4 calculation, a substantial baseline age imbalance between fibrosis groups is structurally expected in observational datasets. Therefore, balancing age between groups is essential when evaluating the relationship between retinopathy and liver fibrosis. Since both retinopathy and liver fibrosis are strongly age-dependent conditions, failure to account for age could produce a spurious association driven primarily by their shared age dependence rather than a direct biological relationship.

In this study, PSM was implemented to enhance between-group comparability and reduce potential confounding effects, and robust standard error estimates were used in post-matching analyses to account for the dependency structure induced by matching. This strategy contributes to a more balanced and cautious interpretation of model-based results, particularly in large observational datasets such as NHANES. The limited consideration of such balancing approaches in previous studies may partially explain the inconsistencies reported in the literature (Rogers, 2023; Salerno et al., 2025).

The LASSO regression method used in the variable selection process is widely recommended as a parsimonious and stable modeling approach for high-dimensional data with multicollinearity. Traditional stepwise regression techniques are prone to model instability and overfitting. In contrast, LASSO regression applies L_1 regularization to shrink coefficients and exclude weaker predictors, enabling the construction of more concise models (Schonlau, 2023; Bangchang, 2024). In the present analysis, LASSO-based variable selection indicated that several clinical and demographic variables provided limited independent contribution in the presence of other covariates and were therefore excluded from the final model. This approach aimed to reduce model complexity and improve interpretability.

In the multivariable logistic regression model constructed using variables selected through LASSO, hepatitis C virus infection, body mass index, and height were identified as factors associated with liver fibrosis, whereas retinopathy was not confirmed as an independent determinant within this modeling framework. This finding suggests that retinopathy may be related to fibrosis indirectly through shared confounding pathways-such as age, metabolic status, and coexisting systemic conditions-rather than acting as a direct risk factor.

The inverse association observed between BMI and liver fibrosis should be interpreted cautiously. In cross-sectional observational datasets such as NHANES, lower BMI values among individuals with advanced liver disease may partly reflect illness-related weight loss, reverse causation, or underlying disease severity, rather than a true protective effect of higher BMI. Such patterns have been widely discussed in the context of the so-called obesity paradox observed in several chronic diseases (Elsabaawy, 2024). The variability in findings reported in the literature should therefore be interpreted in light of differences in variable selection strategies and heterogeneity in the covariate sets included in analytical models (Feng et al., 2021; Guo et al., 2023; Zhang et al., 2023).

One of the key methodological contributions of this study is the demonstration that associations identified in large observational health datasets may be highly sensitive to analytical design choices. By applying propensity score matching to improve covariate balance and LASSO-based regularization to guide variable selection within the same dataset, this study illustrates how different modeling strategies can substantially influence the magnitude and statistical significance of observed relationships. These findings underscore the importance of transparent and carefully justified analytical frameworks when interpreting epidemiological associations derived from complex real-world datasets such as NHANES.

The strengths of this study include the use of a large population-based NHANES sample, the implementation of propensity score matching to enhance covariate balance, and the application of LASSO-based variable selection to construct a parsimonious multivariable model. By integrating these complementary analytical approaches within the same dataset, the study provides a methodological framework for evaluating how modeling

strategies may influence epidemiological interpretations in large observational health databases.

Several limitations should also be considered. First, due to the cross-sectional design of the study, the findings cannot be interpreted as evidence of causal relationships. Second, PSM can only balance measured confounders, and unmeasured confounding may remain a source of residual bias. In addition, although NHANES provides nationally representative data, the complex survey design and sampling weights were not incorporated into the present modeling framework, which may limit the direct generalizability of the findings to the broader population.

Future studies may benefit from prospective designs and sensitivity analyses to more comprehensively assess the potential impact of unmeasured confounding. In addition, replication of these findings in independent populations and different epidemiological settings would help evaluate the robustness and generalizability of the observed associations. Moreover, explainable machine learning approaches may provide an alternative methodological perspective by enabling the joint evaluation of variable selection and model performance in similar epidemiological research questions. Future research may further evaluate the robustness of these findings using alternative causal inference frameworks, such as inverse probability weighting, doubly robust estimation, or targeted maximum likelihood estimation. In addition, integrating explainable machine learning approaches with traditional epidemiological modeling may provide a promising direction for improving both interpretability and predictive performance in large observational health datasets.

From a methodological perspective, this study contributes to the growing literature emphasizing the importance of analytical transparency in observational health research. By integrating propensity score matching with LASSO-based variable selection within the same analytical framework, the study demonstrates how confounding-control strategies and feature-selection methods can substantially influence statistical inference. These findings highlight the need for carefully justified modeling strategies when analyzing large epidemiological datasets and may guide future studies aiming to combine causal inference approaches with modern regularization techniques.

5. Conclusion

In conclusion, this study demonstrates that the statistical methods used to evaluate the association between retinopathy and liver fibrosis may have a meaningful influence on the resulting estimates. Rather than informing clinical decision-making, the findings emphasize the importance of confounding control and modeling strategies in large-scale observational datasets and contribute to ongoing methodological discussions in epidemiological research.

Author Contributions

The percentages of the authors' contributions are presented below. All authors reviewed and approved the final version of the manuscript.

	F.CT.	B.V.
C	70	30
D	65	35
S	60	40
DCP	55	45
DAI	70	30
L	60	40
W	75	25
CR	55	45
SR	70	30
PM	65	35

C= concept, D= design, S= supervision, DCP= data collection and/or processing, DAI= data analysis and/or interpretation, L= literature search, W= writing, CR= critical review, SR= submission and revision, PM= project management.

Conflict of Interest

The authors declared that there is no conflict of interest.

Ethical Consideration

Ethics committee approval was not required for this study because it was conducted using publicly available, de-identified secondary data. The data were obtained from the National Health and Nutrition Examination Survey (NHANES), which is conducted in accordance with ethical standards and approved by the National Center for Health Statistics Research Ethics Review Board. No identifiable human data were used in this analysis.

References

Asero, C., Giandalia, A., Cacciola, I., Morace, C., Lorello, G., Caspanello, A. R., Alibrandi, A., Squadrito, G., & Russo, G. T. (2023). High Prevalence of Severe Hepatic Fibrosis in Type 2 Diabetic Outpatients Screened for Non-Alcoholic Fatty Liver Disease. J Clin Med, 12(8). https://doi.org/10.3390/jcm12082858

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate behavioral research, 46(3), 399-424.

Bangchang, K. N. (2024). Application of Bayesian variable selection in logistic regression model. AIMS Mathematics, 9(5), 13336-13345. https://doi.org/10.3934/math.2024650

Bo, J., & Yang, M. (2025). Dose-response relationship between physical activity and visceral fat mass: a cross-sectional study based on NHANES 2011-2018. BMC Public Health, 25(1), 3113. https://doi.org/10.1186/s12889-025-24393-6

Burgos-Ochoa, L., & Clouth, F. J. (2025). Causal Inference and Survey Data in Paediatric Epidemiology: Generalising Treatment Effects From Observational Data. Paediatr Perinat Epidemiol. https://doi.org/10.1111/ppe.70042

Cheang, I., Zhu, X., Zhu, Q., Li, M., Liao, S., Zuo, Z., Yao, W., Zhou, Y., Zhang, H., & Li, X. (2022). Inverse association between blood

ethylene oxide levels and obesity in the general population: NHANES 2013-2016. Front Endocrinol (Lausanne), 13, 926971. https://doi.org/10.3389/fendo.2022.926971

Chhabra, S., Singh, S. P., Singh, A., Mehta, V., Kaur, A., Bansal, N., & Sood, A. (2022). Diabetes Mellitus Increases the Risk of Significant Hepatic Fibrosis in Patients With Non-alcoholic Fatty Liver Disease. J Clin Exp Hepatol, 12(2), 409-416. https://doi.org/10.1016/j.jceh.2021.07.001

Deravi, N., Dehghani Firouzabadi, F., Moosaie, F., Asadigandomani, H., Arab Bafrani, M., Yoosefi, N., Poopak, A., Dehghani Firouzabadi, M., Poudineh, M., Rabizadeh, S., Kamel, I., Nakhjavani, M., & Esteghamati, A. (2023). Non-alcoholic fatty liver disease and incidence of microvascular complications of diabetes in patients with type 2 diabetes: a prospective cohort study. Front Endocrinol (Lausanne), 14, 1147458. https://doi.org/10.3389/fendo.2023.1147458

Ellis, G. M., & Souza, P. E. (2021). Using Machine Learning and the National Health and Nutrition Examination Survey to Classify Individuals With Hearing Loss. Front Digit Health, 3, 723533. https://doi.org/10.3389/fdgth.2021.723533

Elsabaawy, M. (2024). Liver at crossroads: unraveling the links between obesity, chronic liver diseases, and the mysterious obesity paradox. *Clinical and experimental medicine*, *24*(1), 240. https://doi.org/10.1007/s10238-024-01493-y

Erman, H., Boyuk, B., Arslan, S., Akin, S., & Keskin, O. (2024). Noninvasive Liver Fibrosis Indices as Indicators of Microvascular and Macrovascular Complications in Type 2 Diabetes. Metab Syndr Relat Disord, 22(8), 619-625. https://doi.org/10.1089/met.2024.0022

Feng, G., Zheng, K. I., Li, Y. Y., Rios, R. S., Zhu, P. W., Pan, X. Y., Li, G., Ma, H. L., Tang, L. J., Byrne, C. D., Targher, G., He, N., Mi, M., Chen, Y. P., & Zheng, M. H. (2021). Machine learning algorithm outperforms fibrosis markers in predicting significant fibrosis in biopsy-confirmed NAFLD. J Hepatobiliary Pancreat Sci, 28(7), 593-603. https://doi.org/10.1002/jhbp.972

Guo, Y., Shen, B., Xue, Y., & Li, Y. (2023). Development and validation of a non-invasive model for predicting significant fibrosis based on patients with nonalcoholic fatty liver disease in the United States. Front Endocrinol (Lausanne), 14, 1207365. https://doi.org/10.3389/fendo.2023.1207365

Issanov, A., Karim, M. E., Aimagambetova, G., & Dummer, T. J. B. (2022). Does Vaccination Protect against Human Papillomavirus-Related Cancers? Preliminary Findings from the United States National Health and Nutrition Examination Survey (2011-2018). Vaccines (Basel), 10(12). https://doi.org/10.3390/vaccines10122113

Jacob, M., Joseph, M., & Idiculla, J. (2023). Non-alcoholic fatty liver disease and diabetic retinopathy: Is there an association? J Family Med Prim Care, 12(9), 2028-2031. https://doi.org/10.4103/jfmpc.jfmpc_2327_22

Jeong, J., & Kim, Y.-M. (2024). Robust estimation of a marginal causal effect on the binary outcome using propensity score matching. Journal of the Korean Data And Information Science Society, 35(1), 161-177. https://doi.org/10.7465/jkdi.2024.35.1.161

Li, J., Xiang, Y., Han, J., Gao, Y., Wang, R., Dong, Z., Chen, H., Gao, R., Liu, C., Teng, G. J., & Qi, X. (2024). Retinopathy as a predictive indicator for significant hepatic fibrosis according to T2DM status: A cross-sectional study based on the national health and nutrition examination survey data. Ann Hepatol, 29(4), 101478. https://doi.org/10.1016/j.aohep.2024.101478

Mantovani, A., Mozrieri, M. L., Aldigeri, R., Palmisano, L., Masulli, M., Bonomo, K., Baroni, M. G., Cossu, E., Cimini, F. A., Cavallo, G., Buzzetti, R., Mignogna, C., Leonetti, F., Bacci, S., Trevisan, R., Pollis, R. M., Cas, A. D., de Kreutzenberg, S. V., & Targher, G. (2024). MASLD, hepatic steatosis and fibrosis are associated with the prevalence of chronic kidney disease and retinopathy in adults with type 1 diabetes mellitus. Diabetes Metab, 50(1), 101497. https://doi.org/10.1016/j.diabet.2023.101497

Rogers, P. (2023). Controlling for Confounding in Complex Survey Machine Learning Models to Assess Drug Safety and Risk. In Machine Learning and Deep Learning in Computational Toxicology (pp. 355-374). Springer.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. Biometrika, 70(1), 41-55.

Salerno, S., Roberts, E. K., Needham, B. L., McCormick, T. H., Li, F., Mukherjee, B., & Shi, X. (2025). What's the Weight? Estimating Controlled Outcome Differences in Complex Surveys for Health Disparities Research. Statistics in medicine, 44(23-24), e70289.

Schonlau, M. (2023). Lasso and Friends. In Applied Statistical Learning: With Case Studies in Stata (pp. 73-95). Springer.

Shaji, N., Singhai, A., Sarawagi, R., Pakhare, A. P., Mishra, V. N., & Joshi, R. (2022). Assessment of Liver Fibrosis Using Non-invasive Screening Tools in Individuals With Diabetes Mellitus and Metabolic Syndrome. Cureus, 14(2), e22682. https://doi.org/10.7759/cureus.22682

Storz, M. A. (2023). Does Self-Perceived Diet Quality Align with Nutrient Intake? A Cross-Sectional Study Using the Food Nutrient Index and Diet Quality Score. Nutrients, 15(12). https://doi.org/10.3390/nu15122720

Woodard, J. S., & Abrams, G. A. (2024). Increased Prevalence of Advanced Metabolic Dysfunction-Associated Steatotic Liver Disease Fibrosis in Type 2 Diabetics Despite Low-Risk Fibrosis-4 Index Scores. Journal of Endocrinology and Metabolism, 14(1), 40-47. https://doi.org/10.14740/jem935

Zhang, Z., Wang, J., Wang, H., Qiu, Y., Zhu, L., Liu, J., Chen, Y., Li, Y., Liu, Y., Chen, Y., Yin, S., Tong, X., Yan, X., Xiong, Y., Yang, Y., Zhang, Q., Li, J., Zhu, C., Wu, C., & Huang, R. (2023). An easy-to-use AIHF-nomogram to predict advanced liver fibrosis in patients with autoimmune hepatitis. Front Immunol, 14, 1130362. https://doi.org/10.3389/fimmu.2023.1130362