Predicting Gallstone Presence Using Non-Imaging Clinical Variables

Abstract
Gallstone disease is a common condition linked to metabolic and inflammatory processes, yet diagnosis relies primarily on ultrasound, which may be difficult to access in some clinical environments. In this study, we built a logistic regression model in R to predict gallstone presence using non-imaging clinical variables. Using real patient data from 319 adults in a Turkish outpatient clinic, we applied VIF screening and stepwise procedures under AIC and BIC criteria to identify the best-performing model. After removing one influential outlier, our final BIC-selected model achieved an AUC of 0.9262 using leave-one-out cross-validation. The model identifies inflammation markers, metabolic conditions, and specific body-composition measures as important predictors of gallstone disease.

1. Background and Significance

Gallstone disease is a common gastrointestinal disorder in which solid stones form in the gallbladder. In developed countries, approximately 10-15% of adults are affected by gallstones, and it is a major driver of digestive system hospitalizations [1]. Currently, gallstones are diagnosed primarily using abdominal ultrasonography. Although noninvasive, ultrasound requires trained specialists, and this equipment is not always available in resource-limited settings [2]. If non-imaging variables (i.e. routine blood tests and body composition measures) could reliably predict gallstone presence, clinicians might be able to triage patients more effectively for earlier intervention and scalable screening.

In this project, we used logistic regression to quantify the associations between non-imaging clinical predictors and gallstone presence. Specifically, our research question is: Among patients in the clinical sample, which non-imaging predictors (metabolic biomarkers, body composition measures, and binary clinical indicators) are significantly associated with gallstone presence, and how large are their effects on the odds of gallstones?

2. Data

2.1 Data Description

Our analysis uses the “Gallstone” dataset from the UCI Machine Learning Repository, originally collected between June 2022 and June 2023 at an internal medicine outpatient clinic in Turkey [3]. After excluding patients with prior gallbladder surgery, the final sample includes 319 adult patients, of whom 161 were diagnosed with gallstones and 158 were gallstone-free, yielding a balanced dataset [4]. The dataset includes a total of 38 non-imaging predictors that fall into four categories (see Table 1). Our response variable is Gallstone, a binary indicator coded as 1 for patients with gallstones and 0 for patients without gallstones.

2.2 Data Cleaning

Our dataset contained no missing values, and all 38 predictors were either binary or continuous. We examined multicollinearity issues by conducting a Variance Inflation Factor (VIF) analysis on all predictor variables. Using a threshold of 10, we identified and removed eight predictors, leaving a final dataset with 30 predictors and 319 observations.

3. Methods and Results

3.1 Model Selection

For model fitting, we renamed some lengthy variable names to abbreviations (see Table 2). To identify the best first-order model, we fit a logistic regression model using all predictors in the final dataset and adopted stepwise regression in both directions with AIC and BIC criteria. Since the two criteria suggested different best models (Model 1 with AIC criterion and Model 2 with BIC criterion), we compared their performance using the leave-one-out cross-validation and computed multiple metrics for binary classifications. Using a classification threshold of 0.5, the two models yielded identical confusion matrices. Model 2 had a higher AUC and was more parsimonious, so we selected Model 2 as the optimal first-order model.

We then assessed models with interaction terms to determine whether they significantly improve model fit. Building upon Model 2, we conducted stepwise regression in both directions with all interaction terms using both AIC and BIC criteria and identified Model 3 (AIC criterion) and Model 4 (BIC criterion). We computed performance metrics for both models and found that Model 3 outperformed all other models (Table 3). However, interpreting coefficients in a model with interaction terms depends on the values of other variables. Since our research question focused on the magnitude of effects, selecting a model with interaction terms may limit our

ability to address our question. Given its adequate performance, we selected Model 2 as the final model.

Table 3. Comparison of first-order (Model 1 and 2) and interaction (Model 3 and 4) models

	Number of Predictors	AUC	Accuracy	Sensitivity	Specificity	Precision	F-measure
Model 1	17	0.8910	0.8119	0.7595	0.8634	0.8451	0.8
Model 2	**10**	0.8918	0.8119	0.7595	0.8634	0.8451	0.8
Model 3	14	**0.9104**	**0.8464**	**0.8101**	**0.8820**	**0.8707**	**0.8393**
Model 4	11	0.9100	0.8401	**0.8101**	0.8696	0.8591	0.8339

3.2 Model Diagnostics

We first examined residual plots for Model 2, identifying four potential outliers (see Figure 1). Then, to check whether these outliers are influential, we computed the delta deviance for each observation (see Figure 2). Observation 154 stood out and appeared to be influential, so we removed it. All other observations were not as influential and were kept. We inspected observation 154 and found that this patient has the lowest Hemoglobin and highest C-reactive protein levels. In our model, gallstone presence is negatively associated with Hemoglobin and positively associated with CRP. This combination predicts a high probability of gallstones, yet this patient was gallstone-free, which explains its elevated delta deviance.

3.3 Results

After removal of the influential outlier identified in the diagnostic analysis, our final model was a BIC-selected first-order logistic regression with ten predictors: Hyperlipidemia, Diabetes Mellitus (DM), Intracellular Water (ICW), HDL cholesterol, C-reactive protein (CRP), liver enzymes ALT and AST, Bone Mass (BM), Hemoglobin (HGB), and Vitamin D (VitD). Removing the outlier improved overall model performance across all classification metrics.

Several predictors show strong and statistically significant associations with gallstone presence. In particular, CRP emerges as one of the strongest risk factors: holding all other predictors constant, a one-unit increase in CRP multiplies the predicted odds of gallstones by approximately 3.87, indicating a substantial association between systemic inflammation and gallstone disease. Similarly, higher levels of ALT are associated with increased odds of gallstones, while AST shows a negative association after adjusting for other variables.

In contrast, Bone Mass and Vitamin D are strongly negatively associated with gallstone presence. For each additional kilogram of bone mass, the predicted odds of gallstones decrease substantially, after adjusting for simultaneous linear changes of all other predictors in the model. Higher Vitamin D levels are also associated with reduced odds of gallstones, suggesting a protective relationship in this clinical sample. Hemoglobin exhibits a negative association as well, indicating lower hemoglobin levels correspond to increased gallstone risk.

Overall, the final BIC first-order logistic model excluding one influential outlier yields high discriminative performance (AUC ≈ 0.93) and identifies a set of interpretable, clinically meaningful non-imaging predictors of gallstone disease (Table 4).

Table 4. Comparison of final model before and after removing outlier

	AUC	Accuracy	Sensitivity	Specificity	Precision	F-measure
Before Removal	0.8910	0.8119	0.7595	0.8634	0.8451	0.8
After Removal	0.9262	0.8428	0.8101	0.875	0.8649	0.8367

4. Discussion and Other Considerations

4.1 Conclusion

Our findings highlight several clinically intuitive patterns. Increases in CRP and ALT, markers of systemic inflammation and hepatocellular injury, are strongly associated with higher odds of gallstones, whereas higher vitamin D levels and greater bone mass are associated with lower odds. Diabetes and hyperlipidemia, two conditions tied to metabolic syndrome, are also important predictors, reinforcing the view that gallstone disease is intertwined with broader metabolic dysregulation [5]. The association of vitamin D aligns with the literature suggesting that vitamin D deficiency may impair gallbladder motility and promote bile stasis [6]. The association between bone mass and gallstones is also consistent with prior work reporting increased gallstone risk among patients with osteoporosis [7]. These associations support the biological plausibility of our model and suggest that gallstone disease reflects a combination of inflammatory, metabolic, and body-composition–related factors.

Taken together, our results suggest that a parsimonious logistic model using non-imaging measures can discriminate reasonably well between patients with and without gallstones in this clinical setting. While ultrasound remains the standard for definitive diagnosis, models of this kind demonstrate that routinely collected laboratory and body-composition measures contain meaningful information about gallstone risk. As a result, such models may serve as a useful preliminary screening or triage tool to help prioritize patients for further imaging, particularly in contexts where imaging resources are constrained.

4.2 Limitations and Future Directions

Our study has several limitations that should be considered. First, the sample may not capture the diversity of patients from other regions, ethnicities, or healthcare systems. As a result, extrapolating our findings to a broader population requires validation across heterogeneous demographic cohorts. In particular, body composition measures may vary depending on the device and clinical conditions. Another limitation is that our analysis is cross-sectional, which allows us to report associations between predictors and gallstone presence but does not infer causality. In addition, our model selection was based on slightly better AUC values for the first-order model, and although interaction models using AIC showed improved cross-validation performance, we chose the simpler model to maintain interpretability. Lastly, removing an influential outlier may have excluded a true but rare clinical case, which could affect our model's generalizability to unusual presentations.

Future work could validate the model on a larger and more diverse patient population to ensure broader application. Interaction effects can also be explored, ideally in collaboration with clinicians to identify biologically plausible interactions. Finally, incorporating additional clinical predictors such as medication history, dietary habits, and lifestyle could potentially improve the model's predictive performance and clinical value.

References

[1] Stinton, L. M., & Shaffer, E. A. (2012). *Epidemiology of gallbladder disease: cholelithiasis and cancer.* Gut and Liver, 6(2), 172–187. https://www.ncbi.nlm.nih.gov/books/NBK459370/.

[2] Cleveland Clinic Journal of Medicine. (2018). *Gallstones: Watch and wait, or intervene?* Cleveland Clinic Journal of Medicine, 85(4), 323–331. https://www.ccjm.org/content/85/4/323

[3] Esen, I., Arslan, H., Aktürk, S., Gülşen, M., Kültekin, N., & Özdemir, O. (2024). Gallstone [Dataset]. UCI Machine Learning Repository. https://doi.org/10.1097/md.0000000000037258.

[4] Esen, İrfan MDa,*; Arslan, Hilalb; Aktürk Esen, Selin MDc; Gülşen, Mervenurd; Kültekin, Nimetd; Özdemir, Oğuzhan MDe. Early prediction of gallstone disease with a machine learning-based method from bioimpedance and laboratory data. Medicine 103(8):p e37258, February 23, 2024. | DOI: 10.1097/MD.0000000000037258

[5] Ratheesh R, Ulrich MT, Ghozy S, Al-Jaboori M, Nayak SS. The association between diabetes and gallstones: a nationwide population-based cohort study. Prz Gastroenterol. 2023;18(3):292-299. doi: 10.5114/pg.2023.131395. Epub 2023 Sep 22. PMID: 37937115; PMCID: PMC10626383.

[6] Shabanzadeh DM, Jørgensen T, Linneberg A, Sørensen LT, Skaaby T. Vitamin D and gallstone disease-A population-based study. Endocrine. 2016 Dec;54(3):818-825. doi: 10.1007/s12020-016-1113-4. Epub 2016 Oct 1. PMID: 27696253.

[7] Klahan S, Kuo CN, Chien SC, Lin YW, Lin CY, Lin CH, Chang WC, Lin CI, Hung KS, Chang WP. Osteoporosis increases subsequent risk of gallstone: a nationwide population-based cohort study in Taiwan. BMC Gastroenterol. 2014 Nov 18;14:192. doi: 10.1186/s12876-014-0192-z. PMID: 25404001; PMCID: PMC4247648.

Appendix

Figure 1. Residual plots to identify outliers

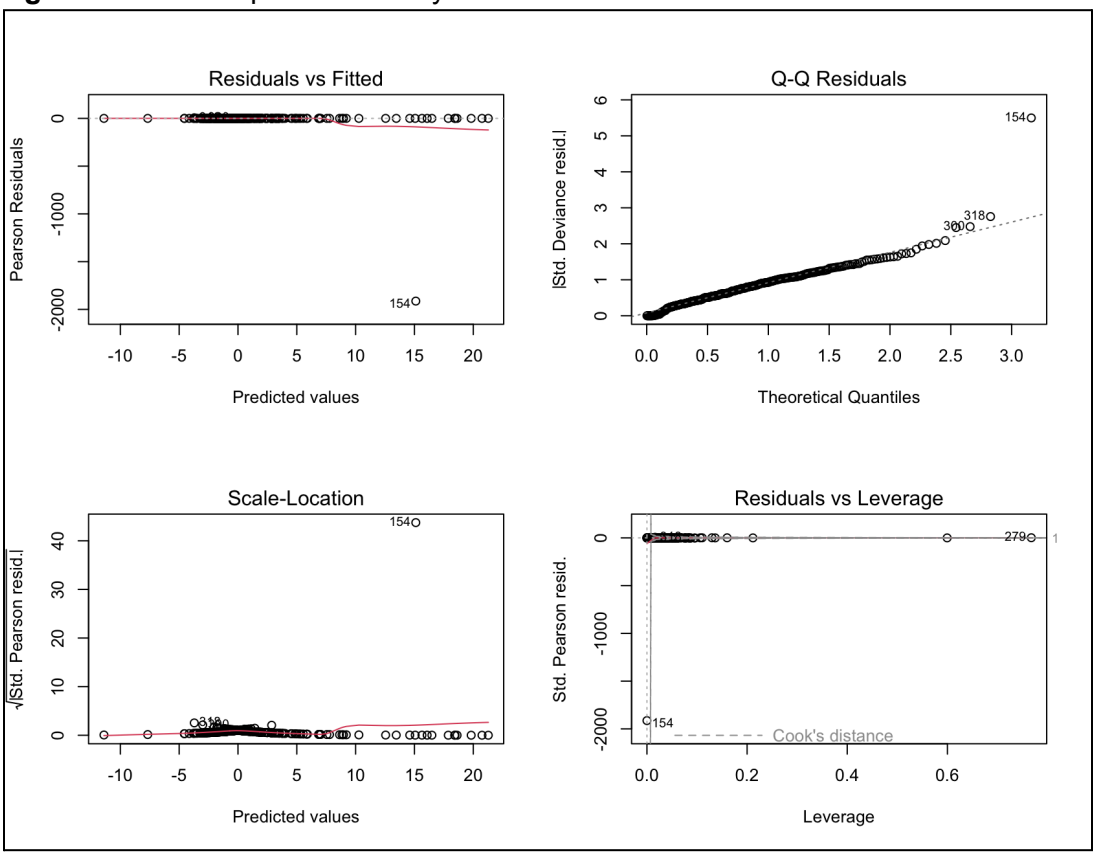

Figure 2. Delta deviance plot to identify influential outliers

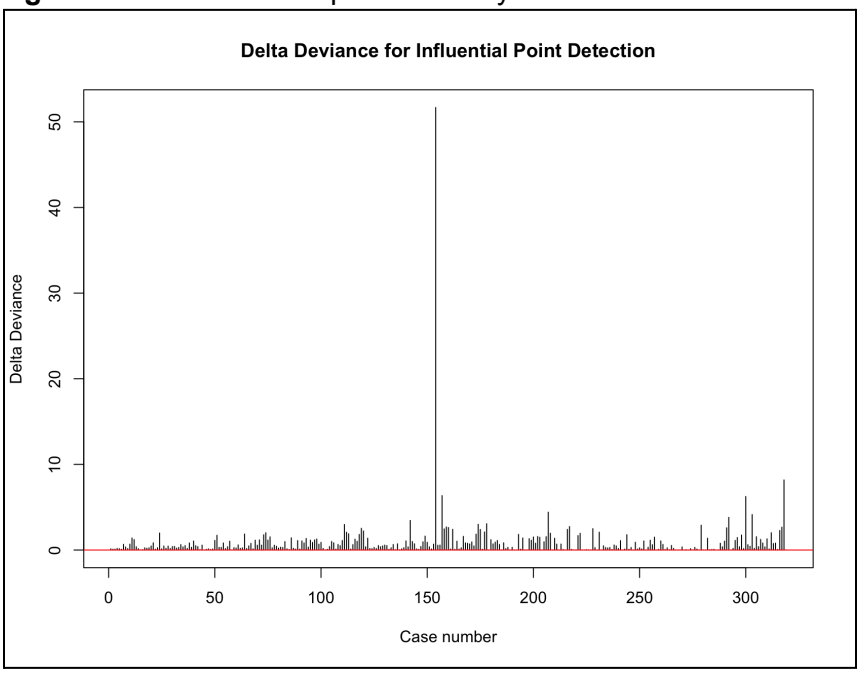

Figure 3. Final Model Output (BIC-selected first-order logistic regression)

```
Call:
glm(formula = Gallstone ~ Hyperlipidemia + DM + ICW + BM + HDL +
    AST + ALT + CRP + HGB + VitD, family = binomial, data = gall2)

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      8.79878    2.33623   3.766 0.000166 ***
Hyperlipidemia  19.24427 1120.04306   0.017 0.986292
DM               1.30849    0.49322   2.653 0.007979 **
ICW              0.71640    0.15725   4.556 5.22e-06 ***
BM              -6.48525    1.48010  -4.382 1.18e-05 ***
HDL              0.01902    0.01398   1.361 0.173578
AST             -0.08784    0.03284  -2.675 0.007473 **
ALT              0.03271    0.01639   1.995 0.046030 *
CRP              1.35209    0.22881   5.909 3.44e-09 ***
HGB             -0.44622    0.13604  -3.280 0.001037 **
VitD            -0.09881    0.01972  -5.010 5.43e-07 ***
---
```

Table 1. Variable Names by Category

Demographics	Body-Composition measures	Metabolic biomarkers	Binary clinical indicators
Age Gender Height Weight	Total Body Water Extracellular Water Intracellular Water Extracellular Fluid Ratio Body Protein Content Bone Mass Total Fat Content Visceral Muscle Area Hepatic Fat Accumulation BMI Lean Mass Visceral Fat Area Muscle Mass Total Body Fat Ratio Visceral Fat Rating	C-Reactive Protein (CRP) Vitamin D Total Cholesterol LDL Cholesterol HDL Cholesterol Triglycerides Glucose Creatinine GFR Liver enzymes (AST, ALT, ALP) Hemoglobin (HGB)	Comorbidity Coronary Artery Disease (CAD) Hypothyroidism Hyperlipidemia Diabetes Mellitus (DM) Obesity indicator

Table 2. Rename Lengthy Variables

Original Name	**Assigned Name (“.” means no change)**	**Variable Type**	**Status after VIF**
Gallstone.Status	Gallstone	Response	Included
Age	.	Predictor	Included
Gender	.	Predictor	Included

Height	.	Predictor	Included
Weight	.	Predictor	Removed
Total Body Water	TBW	Predictor	Included
Extracellular Water	ECW	Predictor	Removed
Intracellular Water	ICW	Predictor	Included
Extracellular Fluid Ratio	ECF_TBW	Predictor	Included
Body Protein Content	Protein	Predictor	Included
Bone Mass	BM	Predictor	Included
Total Fat Content	TFC	Predictor	Included
Visceral Muscle Area	VMA	Predictor	Included
Hepatic Fat Accumulation	HFA	Predictor	Included
BMI	.	Predictor	Removed
Lean Mass	LM	Predictor	Removed
Visceral Fat Area	VFA	Predictor	Removed
Muscle Mass	MM	Predictor	Removed
Total Body Fat Ratio	TBFR	Predictor	Removed
Visceral Fat Rating	VFR	Predictor	Removed
C Reactive Protein	CRP	Predictor	Included
Vitamin D	VitD	Predictor	Included
Total Cholesterol	TC	Predictor	Included
Low Density Lipoprotein Cholesterol	LDL	Predictor	Included
High Density Lipoprotein Cholesterol	HDL	Predictor	Included
Triglycerides	.	Predictor	Included

Glucose	.	Predictor	Included
Creatinine	.	Predictor	Included
Glomerular Filtration Rate	GFR	Predictor	Included
Liver enzymes (Aspartat Aminotransferaz, Alanin Aminotransferaz, Alkaline Phosphatase)	AST, ALT, ALP	Predictor	Included
Hemoglobin	HGB	Predictor	Included
Comorbidity	.	Predictor	Included
Coronary Artery Disease	CAD	Predictor	Included
Hypothyroidism	.	Predictor	Included
Hyperlipidemia	.	Predictor	Included
Diabetes Mellitus	DM	Predictor	Included
Obesity	.	Predictor	Included