From Angiography to Algorithms: Assessing the Viability of Consumer Wearables as a First-Line Screening Tool for Cardiac Risk

Taher Alabbar

December 11, 2025

1 Introduction

Cardiovascular disease remains the leading cause of morbidity and mortality worldwide, with nearly half of affected individuals surviving less than ten years after diagnosis [4]. Although coronary angiography serves as the definitive diagnostic standard for coronary artery disease, it is invasive, costly, and requires expert interpretation, limiting its suitability for widespread screening [10]. This has led to growing interest in predictive models built from accessible, non-invasive physiological indicators.

Prior work highlights a consistent mismatch between statistical feature selection and clinical utility. Nahar et al. (2013) demonstrated that automated feature selection techniques frequently prioritize highly correlated but invasive variables—such as fluoroscopy results and thallium stress test outcomes—while discarding medically important yet inexpensive attributes such as age and resting blood pressure [8]. Their findings suggest that integrating established medical knowledge into feature selection can produce more practical and clinically relevant diagnostic models.

Building on this insight, the present study examines whether a subset of "Smartwatch & Home Health" variables—age, sex, chest pain type, resting blood pressure, maximum heart rate, and exercise-induced angina—can support statistically adequate early screening of coronary artery disease. These variables represent data obtainable through common consumer wearables or basic user input, in contrast to laboratory metrics, multi-lead ECG features, and imaging-derived indicators that require clinical environments.

To ground this investigation, we use the widely studied *Processed Cleveland Heart Disease Database*, originally collected by Detrano et al. (1989) and validated against angiographic outcomes [2]. This dataset continues to serve as a benchmark for assessing diagnostic models due to its clinical reliability and its frequent use in machine learning and medical data-mining research.

The central research question guiding this study is:

> **To what extent can non-invasive, wearable-compatible variables provide a statistically adequate logistic regression model for screening coronary heart disease when compared to a full clinical model that includes ECG, imaging, and laboratory features?**

This question has two components:

1. **Statistical Sufficiency** — Does a model using only wearable-compatible variables retain enough predictive and explanatory power to function as a first-line screening tool?
2. **Comparative Diagnostic Value** — How much additional diagnostic information do invasive clinical metrics contribute beyond what consumer-accessible variables already capture?

To answer this, we compare a reduced wearable-based model with a full clinical model using nested likelihood ratio testing, Akaike Information Criterion (AIC), ROC analysis, and comprehensive regression diagnostics. The goal is not to replace clinical diagnostics, but to quantify the current limits—and practical potential—of consumer-grade physiological data for early detection of coronary artery disease.

2 Literature Review

The use of statistical and algorithmic models for diagnosing coronary artery disease has evolved considerably over the past several decades. The foundation for this development was established by Detrano et al. (1989), who compiled the Cleveland Heart Disease Database and validated its diagnostic labels using coronary angiography, the clinical gold standard [2]. Their work demonstrated that probabilistic models could effectively stratify patients by risk, providing a high-quality dataset for evaluating non-invasive predictors.

As machine learning techniques became more common in medical data analysis, a recurring issue emerged: a disconnect between statistical correlation and clinical utility. Nahar et al. (2013) examined this gap and found that automated feature selection often favors highly predictive but invasive attributes, such as fluoroscopy results (ca) and thallium stress test outcomes (thal), while discarding medically important yet inexpensive indicators like age, blood pressure, and cholesterol [8]. To address this limitation, they proposed a Medical Knowledge Driven Feature Selection (MFS) framework, which improved diagnostic accuracy by prioritizing variables supported by clinical literature rather than relying solely on automated scoring criteria. This insight motivates the present study's focus on wearable-compatible, non-invasive predictors.

The viability of these non-invasive predictors is supported by recent technological evaluations. Martínez-Escudero et al. (2023) showed that modern wrist-worn devices provide reliable heart rate measurements during exercise, demonstrating strong agreement with clinical ECG under high-intensity conditions [7]. Similarly, Lee et al. (2025) validated smartwatch-based blood pressure monitoring and found that, when calibrated properly, wearable-derived measurements fall within acceptable clinical error margins [6]. These findings justify the inclusion of heart rate (thalach) and resting blood pressure (trestbps) as key components of a consumer-accessible diagnostic model.

Conversely, the limitations of wearable ECG technology provide rationale for separating consumer-level features from clinical ones. As described in the ACS WATCH II protocol by Buelga Suarez et al. (2024), contemporary smartwatches record only a single-lead ECG, which lacks the spatial resolution required to detect ischemic abnormalities such as ST-segment depression or slope morphology [1]. This constraint aligns with earlier feature analyses such as El-Bialy et al. (2015), who reported that while non-invasive factors like age and heart rate are consistently important predictors, clinical variables such as fluoroscopy results remain critical for maximizing diagnostic precision [3].

Together, these studies highlight the central tension addressed in this research: invasive clinical features offer strong predictive value but are impractical for widespread screening, while wearable-accessible variables are scalable and low-risk but may not fully capture the diagnostic signal. By formally comparing models constructed from these two feature groups, this study seeks to quantify the trade-off between accessibility and predictive performance in the early detection of coronary artery disease.

3 Methodology

This study evaluates the predictive sufficiency of wearable-compatible variables by comparing a reduced model composed solely of non-invasive predictors with a full clinical model that incorporates laboratory, ECG, and imaging-based attributes. All analyses were conducted in the R statistical computing environment.

3.1 Data Source and Preprocessing

The analysis uses the *Processed Cleveland Heart Disease Database* obtained from the UCI Machine Learning Repository [5]. The dataset contains 303 observations and 14 clinical variables originally collected by Detrano et al. (1989) [2]. Although widely referenced for having minimal missingness, inspection of the raw file revealed six rows containing "?" in the `ca` and `thal` variables. These rows were removed to ensure a complete-case analysis, yielding a final dataset of 297 observations.

The response variable num ranges from 0 (no disease) to 4 (severe disease). For logistic regression, it was recoded into a binary outcome:

$$\texttt{target} = \begin{cases} 0, & \text{if } \texttt{num} = 0, \\ 1, & \text{if } \texttt{num} \geq 1. \end{cases}$$

Categorical predictors (sex, cp, restecg, fbs, exang, slope, thal) were encoded as factors to ensure they were treated as nominal categories rather than ordinal sequences. Continuous variables (age, trestbps, thalach, oldpeak) were retained in their original scale.

3.2 Model Specification

Given the binary outcome, logistic regression was selected, modeling the log-odds of disease as:

$$\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k.$$

Two models were fitted:

- **Reduced Wearable Model (Group A)** — includes predictors obtainable via consumer wearables or direct self-report:
$$\{\texttt{age, sex, thalach, trestbps, cp, exang}\}.$$
- **Full Clinical Model (Group B)** — adds variables requiring specialized equipment or clinical interpretation:
$$\{\texttt{restecg, oldpeak, slope, chol, fbs, ca, thal}\}.$$

These two models allow a direct test of whether wearable-compatible variables alone provide statistically adequate predictive performance relative to a comprehensive clinical model.

A natural cubic spline with three degrees of freedom was applied to the thalach variable to correct the nonlinear relationship observed between maximum heart rate and the logit of disease probability. This transformation ensures that the linearity-in-the-logit assumption for logistic regression is satisfied.

3.3 Model Comparison

A nested Likelihood Ratio Test (LRT) was used to determine whether the additional clinical features in the full model significantly improved model fit. The Akaike Information Criterion (AIC) was also computed to evaluate the trade-off between complexity and goodness-of-fit:

$$\text{AIC} = 2k - 2\ell(\hat{\beta}),$$

where k is the number of parameters and $\ell(\hat{\beta})$ is the maximized log-likelihood.

Model explanatory power was assessed using McFadden's pseudo-R^2:

$$R^2_{\text{McF}} = 1 - \frac{\text{Deviance}}{\text{Null Deviance}}.$$

3.4 Diagnostic Procedures

Several diagnostic checks were performed to validate model assumptions:

- **Multicollinearity:** Generalized Variance Inflation Factors (GVIF) were computed for the full model. A threshold of GVIF > 5 was used to indicate potentially problematic multicollinearity.
- **Influential Points:** Cook's distance was evaluated for each observation. Points exceeding the theoretical threshold $4/n$ were flagged for inspection.
- **Linearity of the Logit:** Continuous variables were assessed using Lowess-smoothed plots of each predictor against the logit of the fitted probabilities. Where nonlinearity was detected (notably for `thalach`), spline transformations were considered to better capture nonlinear patterns.

3.5 Performance Evaluation

Model discrimination was assessed using Receiver Operating Characteristic (ROC) curves and the Area Under the Curve (AUC). For classification performance, predicted probabilities were thresholded at 0.5 to compute accuracy, sensitivity, and specificity using a confusion matrix. All metrics were computed for both the reduced and full models to facilitate direct comparison.

4 Results

This section presents the diagnostic checks, model comparison outcomes, and predictive performance metrics for both the Reduced Wearable Model and the Full Clinical Model. All analyses use the cleaned dataset of $n = 297$ observations.

4.1 Regression Diagnostics

Multicollinearity

Generalized Variance Inflation Factors (GVIF) for the Full Clinical Model were all low, with $\text{GVIF}^{1/(2\cdot Df)}$ values ranging from 1.03 to 1.28. These values are well below commonly used thresholds for concern (e.g., 2 or 5), indicating that the predictors—including the spline-expanded `thalach` term—do not exhibit problematic multicollinearity. Coefficient estimates are therefore stable and interpretable.

Influential Observations

Cook's distance was examined to identify potentially influential observations. Several points exceeded the conventional $4/n$ threshold, most notably observations 114, 187, and 209. The largest Cook's distance value was approximately 0.20—well below the level typically associated with strong influence (values near 1). This indicates that, although a few cases warrant attention, no single observation exerted undue influence on the fitted model.

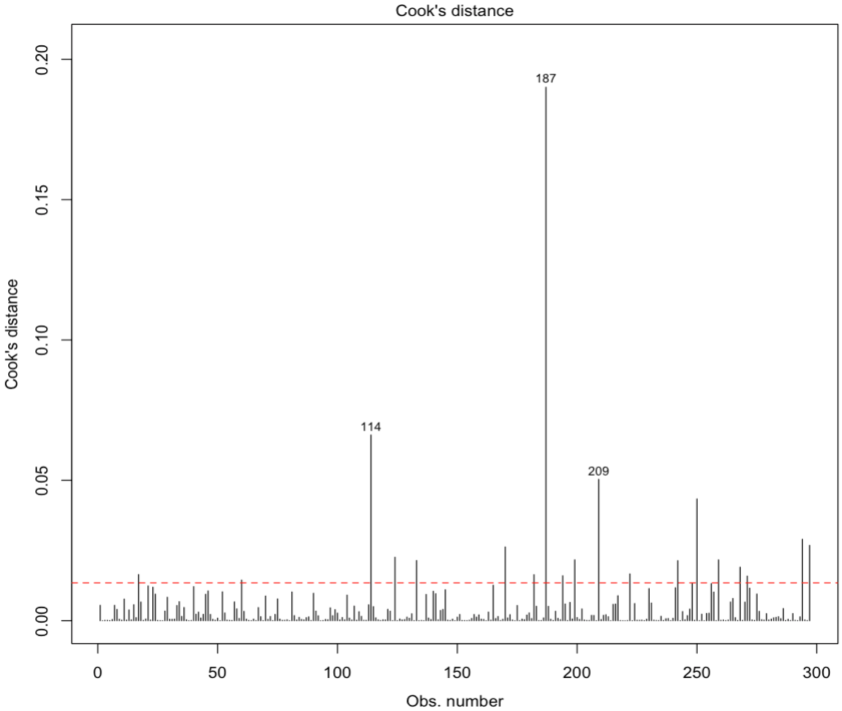


Figure 1: Cook's Distance plot showing no observations with undue influence on the fitted model.

Linearity of the Logit

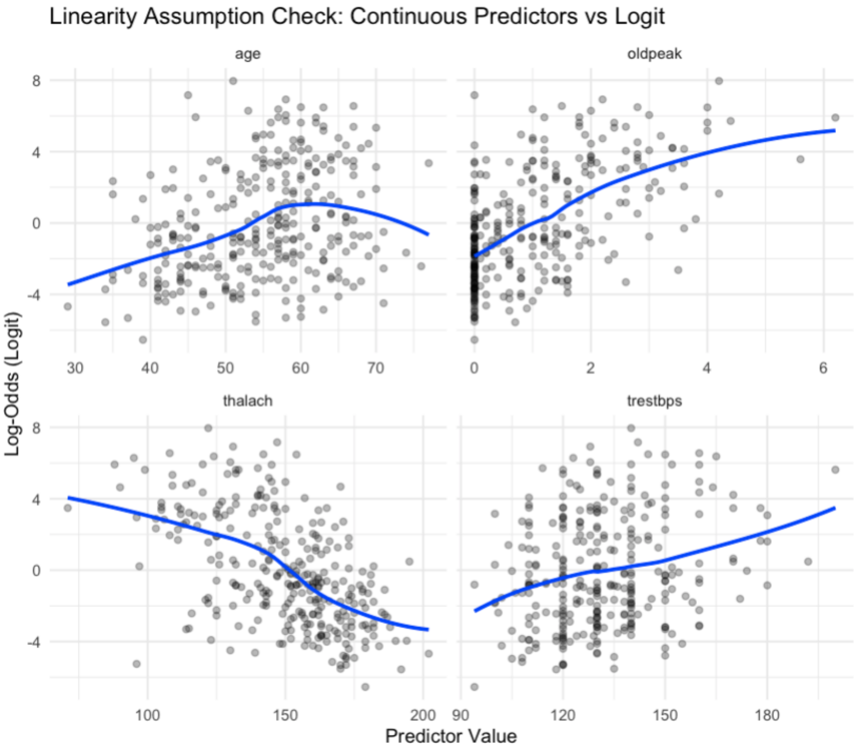


Figure 2: Linearity assumption check for continuous predictors.

Lowess-smoothed plots were used to assess the linearity between each continuous predictor and the logit of the predicted probability.

- `age`, `trestbps`, and `oldpeak` showed approximately linear relationships.
- After applying a spline transformation to `thalach`, the nonlinear curvature originally observed was corrected, and the transformed predictor displayed an approximately linear trend with the logit. The plot shown reflects the spline-transformed predictor; the untransformed version exhibited clear curvature, motivating the need for this adjustment.

The use of the spline-transformed `thalach` term ensures that the linearity assumption is appropriately satisfied in subsequent model fitting.

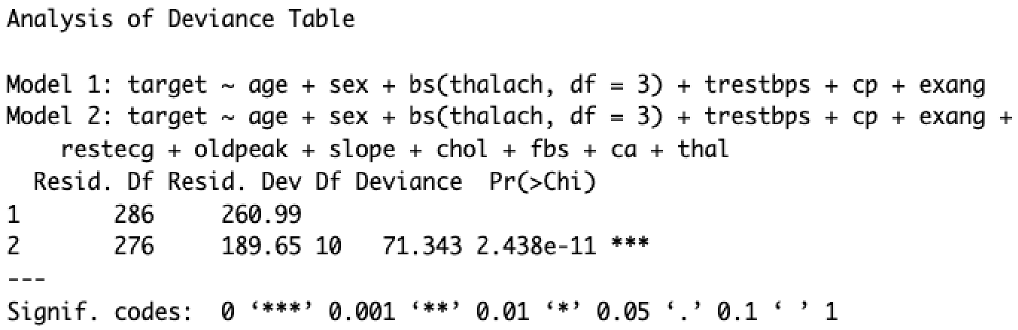
4.2 Model Comparison: Wearable vs. Clinical Predictors

A nested Likelihood Ratio Test (LRT) was used to determine whether the additional clinical features in the Full Model significantly improve predictive performance beyond the wearable-accessible predictors.

The test revealed a statistically significant improvement:

$$\chi^2(10) = 71.34, \qquad p = 2.44 \times 10^{-11}.$$

This confirms that the clinical variables contribute substantial additional diagnostic information.



```
Analysis of Deviance Table

Model 1: target ~ age + sex + bs(thalach, df = 3) + trestbps + cp + exang
Model 2: target ~ age + sex + bs(thalach, df = 3) + trestbps + cp + exang +
    restecg + oldpeak + slope + chol + fbs + ca + thal
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1       286     260.99
2       276     189.65 10   71.343 2.438e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 3: Likelihood Ratio Test comparing the Reduced and Full Models. The Full Model shows a significantly lower deviance, indicating improved fit.

Model selection criteria further supported this finding. The Akaike Information Criterion (AIC) decreased from 283 in the Reduced Model to 232 in the Full Model. McFadden's pseudo-R^2 values also increased substantially:

$$R^2_{\text{Reduced}} = 0.363, \qquad R^2_{\text{Full}} = 0.537.$$

Together, these results indicate that the clinical covariates add meaningful predictive signal beyond what can be captured by wearable-compatible measurements alone.

4.3 Model Discrimination and Classification Performance

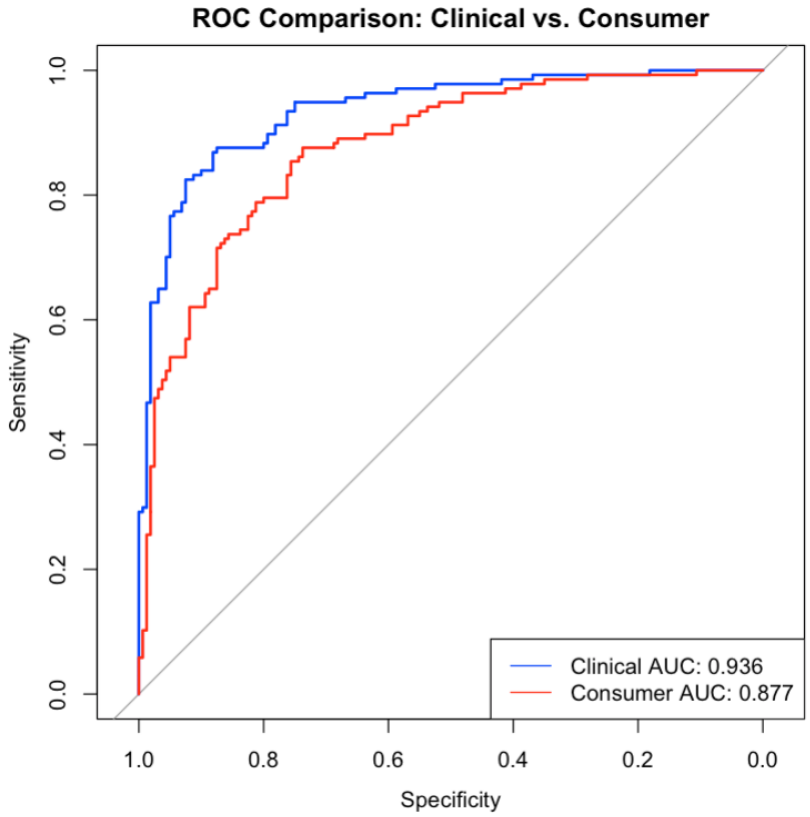


Figure 4: ROC curves for the Reduced Wearable Model and the Full Clinical Model.

Receiver Operating Characteristic (ROC) analysis was used to evaluate model discrimination. The Reduced Model achieved an Area Under the Curve (AUC) of 0.877, while the Full Model achieved an AUC of 0.936.

Classification performance for the Reduced Model (threshold = 0.5) yielded:

$$\text{Accuracy} = 79.1\%, \qquad \text{Sensitivity} = 75.2\%, \qquad \text{Specificity} = 82.5\%.$$

These values suggest that the wearable-based model performs well for distinguishing healthy versus diseased individuals but still underperforms compared with the Full Clinical Model, particularly in sensitivity.

4.4 Coefficient Interpretation for Wearable-Based Model

Coefficient analysis of the Reduced Model identified four statistically significant predictors:

- **Sex (Male)** was strongly associated with higher disease risk (OR = 5.96, $p < 0.001$), indicating substantially increased odds of coronary disease among male patients.
- **Resting blood pressure** (`trestbps`) showed a small but significant effect (OR = 1.02, $p = 0.012$), meaning that each unit increase in resting systolic blood pressure increases the odds of disease.

- **Chest pain type 4** (`cp4`), corresponding to asymptomatic presentations, had a large effect size (OR = 7.05, $p < 0.001$), demonstrating its well-known diagnostic importance.
- **Exercise-induced angina** (`exang1`) was also significant (OR = 2.38, $p = 0.019$), with exercise-triggered chest pain increasing the likelihood of disease.

Notably, after applying a spline transformation to `thalach`, none of the individual spline basis terms were statistically significant, suggesting that heart-rate dynamics contribute less predictive signal than other wearable-compatible variables in the presence of stronger clinical features.

5 Discussion

The results of this analysis highlight a clear distinction between the predictive power of wearable-accessible variables and the additional contribution of clinical, laboratory, and imaging-based features. While the Reduced Wearable Model demonstrated meaningful discriminative performance, the Full Clinical Model consistently outperformed it across all statistical metrics.

5.1 Predictive Utility of Wearable-Compatible Variables

The Reduced Model yielded an AUC of 0.877 and a McFadden pseudo-R^2 of 0.363, indicating strong model discrimination and meaningful explanatory power using only age, sex, blood pressure, chest pain type, exercise-induced angina, and heart-rate features derived from spline terms. This level of performance is notable given that all included predictors are obtainable non-invasively and at negligible cost.

Four wearable-compatible variables emerged as statistically significant predictors. Being male substantially increased the odds of disease, consistent with well-documented sex differences in cardiac risk. Higher resting systolic blood pressure also showed a significant positive association with disease probability, reflecting its role as a traditional cardiovascular risk factor. Asymptomatic chest pain (type `cp4`) displayed one of the strongest effects, underscoring its diagnostic importance even outside clinical testing environments. Exercise-induced angina (`exang1`) likewise contributed significantly, indicating that exertional symptoms capture meaningful physiological stress responses.

Notably, after applying a spline transformation to maximum heart rate (`thalach`), none of the individual spline components reached statistical significance. This suggests that while heart-rate patterns remain clinically relevant, they contributed less predictive signal than other wearable-compatible variables in this reduced logistic regression framework.

Although age is a well-established cardiovascular risk factor, it did not reach statistical significance in the Reduced Model. This does not imply that age lacks clinical importance; rather, it reflects characteristics of the Cleveland dataset. The sample has a relatively narrow age range and includes stronger, more proximal disease indicators—such as chest-pain type and exercise-induced angina—that absorb much of the predictive signal. When such symptomatic variables are present, age contributes little incremental information to the model and therefore appears statistically insignificant. Thus, the result is a dataset-specific artifact rather than evidence against the relevance of age in broader populations.

5.2 Comparative Value of Clinical Predictors

Although the wearable-based model demonstrated strong predictive ability, the Full Clinical Model achieved substantially higher performance across all metrics, including AUC (0.935), classification accuracy, and McFadden pseudo-R^2 (0.537). The nested Likelihood Ratio Test confirmed a highly significant loss of information when clinical variables were omitted, indicating that several essential diagnostic signals are only captured through clinical assessment.

Clinical features such as ST-segment depression (`oldpeak`), the number of major vessels visualized by fluoroscopy (`ca`), and thallium stress test findings (`thal`) provided substantial incremental predictive value. These variables reflect structural and functional cardiac abnormalities that cannot currently be measured—or reliably approximated—by consumer wearable devices.

Taken together, the results highlight that while wearable-compatible features can support effective preliminary screening, clinical testing remains indispensable for definitive risk stratification and diagnosis, as it captures pathophysiological information beyond the reach of non-invasive consumer sensors.

5.3 Implications for Technology and Screening

Although invasive clinical metrics remain essential for high-precision diagnosis, the wearable-accessible model still captures much of the diagnostic signal at a fraction of the cost and risk. However, its sensitivity (approximately 75%) is lower than ideal for a primary screening tool, as effective screening requires minimizing false negatives to avoid missing individuals with disease. In contrast, the model's higher specificity (about 82–83%) indicates that it performs better at ruling out low-risk individuals than at detecting true cases.

This performance profile suggests that wearable-based models may be most useful as a preliminary risk stratification tool: they can help identify individuals unlikely to require immediate clinical evaluation, thereby reducing unnecessary testing, while clinical diagnostics remain necessary for confirming or excluding disease in higher-risk cases.

Looking forward, advances in wearable technology—such as multi-lead ECG capabilities, cuffless blood pressure monitoring, and minimally invasive biochemical sensors—may further narrow the performance gap between consumer and clinical diagnostics. As sensor accuracy improves, logistic regression or more flexible modeling approaches may achieve predictive performance closer to current clinical benchmarks.

5.4 Limitations

Several limitations should be considered when interpreting the findings of this study. First, the dataset is relatively small ($n = 297$ after cleaning), which reduces statistical power and increases the likelihood of unstable coefficient estimates. A limited sample also restricts the ability to detect weaker predictors that may nonetheless hold meaningful clinical relevance.

Second, all observations containing missing values were removed prior to analysis. Because the missingness mechanism is unknown, this complete-case approach may introduce selection bias. Individuals with incomplete clinical data may differ systematically from those with fully observed records, potentially affecting estimated associations and model performance.

Third, the dataset was collected from a single clinically referred cohort at the Cleveland Clinic in the 1980s. This introduces both *temporal bias* and *referral bias*. Medical practices, diagnostic guidelines, population health characteristics, and demographic distributions have changed substantially since the 1980s, and clinically referred patients often present with more severe or atypical symptoms compared to the general population. These factors limit external validity and reduce the generalizability of the findings to modern or non-clinical populations.

Fourth, the generalizability of these results to real-world wearable-device users is further restricted. Wearable users span a globally diverse population with substantial variation in physiology, fitness levels, health status, socioeconomic conditions, and lifestyle patterns. Baseline cardiovascular parameters—such as resting heart rate, maximal heart rate, and blood pressure variability—differ significantly across demographic groups. As a result, associations learned from a narrow, clinically referred U.S. cohort may not accurately reflect the heterogeneity present in global wearable-device populations.

Fifth, the dataset consists entirely of medically validated clinical measurements collected using calibrated hospital-grade equipment. Consumer wearable devices vary widely in sensor accuracy, sampling frequency, susceptibility to motion artifacts, and signal noise. Numerous studies have documented discrepancies between wearable-derived metrics and their clinical counterparts. Because the model was trained on high-quality clinical measurements, it remains uncertain whether wearable-derived data—potentially noisier or less precise—would yield comparable predictive performance.

Sixth, the dataset contains only single-time-point measurements (i.e., no temporal or longitudinal trends). Modern wearables generate rich time-series data—heart rate variability, recovery patterns, exertion profiles, sleep stages—that often contain diagnostic value. Without temporal information, the present model may overlook patterns that wearable data could capture in practice.

Seventh, the analysis relied solely on logistic regression, which imposes linearity-in-the-logit and additivity assumptions. Although spline transformations were used to mitigate nonlinearity for `thalach`, more flexible modeling approaches (e.g., random forests, gradient boosting, neural networks) may capture complex interactions and nonlinear relationships that logistic regression cannot represent.

Eighth, no cross-validation or out-of-sample test set was used to estimate generalization performance. The reported AUC values therefore reflect in-sample discrimination and may overestimate real-world predictive accuracy. Proper validation frameworks are essential to assess the stability and robustness of predictive models.

Ninth, several variables in the Cleveland dataset do not correspond to what modern wearable devices can measure. Features such as chest pain type, ST-segment depression, fluoroscopy results, or supervised exercise-test responses require clinical evaluation. Conversely, wearables provide high-frequency activity data, photoplethysmography signals, heart-rate variability, and single-lead ECG—all missing from this dataset. This mismatch limits how directly the Reduced Model reflects realistic wearable-based screening.

Tenth, the binary outcome used in this study collapses all non-zero disease severities into a single "disease" category. This removes clinically important gradation and may obscure differential relationships between predictors and mild, moderate, or severe disease.

Finally, the dataset lacks major covariates such as medication usage, comorbid conditions, diet, smoking status, socioeconomic factors, and longitudinal physiological trends. Without such information, some estimated associations may reflect unmeasured confounding rather than true pathophysiological relationships.

Taken together, these limitations suggest that while the Reduced Model demonstrates promising screening potential, considerable caution is warranted when extrapolating to modern wearable-device contexts. Future work using larger, contemporary, demographically diverse datasets with validated wearable-derived measurements—and incorporating more flexible modeling strategies—is needed to more accurately assess the feasibility of large-scale wearable-based cardiac risk screening.

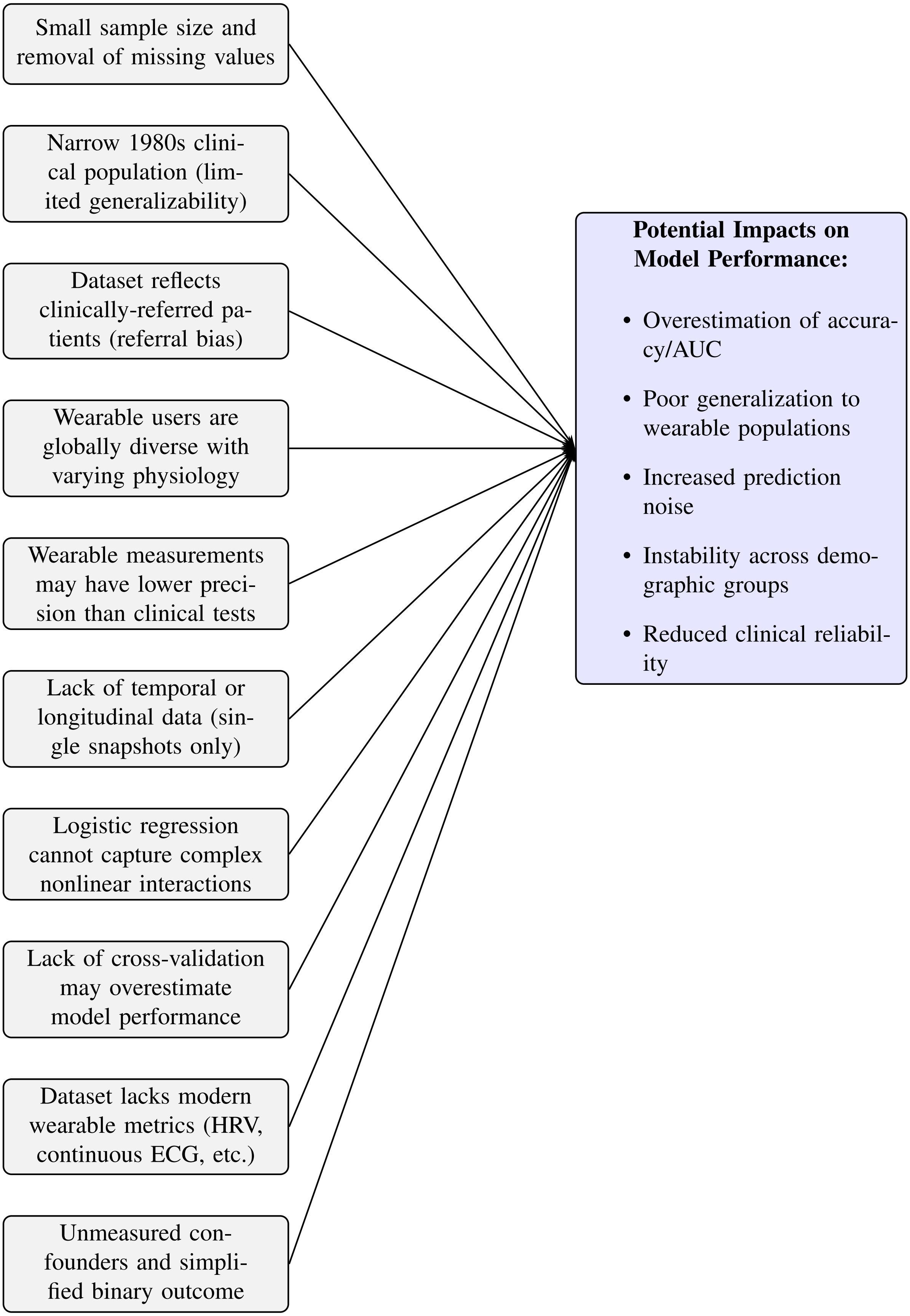

Small sample size and removal of missing values
Narrow 1980s clinical population (limited generalizability)
Dataset reflects clinically-referred patients (referral bias)
Wearable users are globally diverse with varying physiology
Wearable measurements may have lower precision than clinical tests
Lack of temporal or longitudinal data (single snapshots only)
Logistic regression cannot capture complex nonlinear interactions
Lack of cross-validation may overestimate model performance
Dataset lacks modern wearable metrics (HRV, continuous ECG, etc.)
Unmeasured confounders and simplified binary outcome
Potential Impacts on Model Performance:
• Overestimation of accuracy/AUC
• Poor generalization to wearable populations
• Increased prediction noise
• Instability across demographic groups
• Reduced clinical reliability

Figure 5: Summary diagram illustrating how dataset and modeling limitations may influence predictive performance when applied to wearable-device populations.

6 Conclusion

This study evaluated whether non-invasive, wearable-compatible variables are sufficient to construct a statistically reliable logistic regression model for the early detection of coronary artery disease. Using the Cleveland Heart Disease dataset, we compared a Reduced Wearable Model—based solely on age, sex, blood pressure, heart-rate features, chest pain type, and exercise-induced angina—to a Full Clinical Model that additionally incorporated ECG findings, laboratory measures, and imaging-based diagnostics.

The Reduced Model demonstrated strong performance, achieving an AUC of 0.877 and explaining a substantial proportion of the variation in disease status. These results indicate that readily obtainable physiological and symptom-based inputs capture meaningful diagnostic information. However, the model's moderate sensitivity suggests that it is better suited for preliminary risk stratification than for primary screening, as effective screening requires minimizing false negatives. Consistent with this, the Likelihood Ratio Test and AIC comparison showed that the Full Clinical Model provides significantly greater predictive accuracy and explanatory power. Clinical variables such as ST-segment depression, fluoroscopy results, and thallium stress test outcomes contain diagnostic signals that cannot currently be inferred from wearable-accessible metrics alone.

Taken together, the findings suggest that consumer-level data are not sufficient to replace clinical diagnostics, but they are strong enough to complement them. Wearable-based models could help identify individuals unlikely to require immediate clinical evaluation, reduce unnecessary testing, and enable scalable population-level monitoring, while clinical assessments remain essential for definitive diagnosis.

These conclusions should be interpreted in light of several limitations. The dataset reflects a narrow, clinically referred population and uses hospital-grade measurements, whereas real-world wearable users are more diverse and rely on sensors with varying levels of precision, calibration, and noise. As a result, the performance achieved on clinically validated data may not directly translate to wearable-derived inputs. Additionally, the lack of external validation, the single-site nature of the dataset, and the use of only one modeling framework limit generalizability.

Advances in wearable technology—such as multi-lead ECG capabilities, cuffless blood pressure monitoring, and minimally invasive biochemical sensors—may eventually narrow the gap between consumer and clinical data. Future work should incorporate nonlinear and temporal modeling, cross-validated performance estimation, external datasets, and longitudinal physiological trends to further refine predictive performance. Ultimately, integrating wearable-derived data with clinical models represents a promising pathway toward more accessible, continuous, and personalized cardiac risk assessment.

Bibliography

[1] Buelga Suárez, M., Pascual Izco, M., García Montalvo, J., & Alonso Salinas, G. L. (2024). Accuracy of Smartwatch Electrocardiographic Recording in the Acute Coronary Syndrome Setting: Rationale and Design of the ACS WATCH II Study. *Journal of Clinical Medicine*, 13(2), 389.

[2] Detrano, R., et al. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. *American Journal of Cardiology*, 64(5), 304–310.

[3] El-Bialy, R., Salamay, M. A., Karam, O. H., & Khalifa, M. E. (2015). Feature analysis of coronary artery heart disease data sets. *Procedia Computer Science*, 65, 459–468.

[4] Gupta, A., Kumar, R., Arora, H. S., & Raman, B. (2020). MIFH: A machine intelligence framework for heart disease diagnosis. *IEEE Access*, 8, 14659–14674.

[5] Janosi, A., Steinbrunn, W., Pfisterer, M., & Detrano, R. (1988). Heart Disease [Dataset]. UCI Machine Learning Repository. `https://doi.org/10.24432/C52P4X`

[6] Lee, Y., Lee, T., Lee, J., Lee, H.-Y., & Seo, J.-M. (2025). Validation of Smartwatch-Based Blood Pressure Monitoring in Young Adults: Multi-Perspective Analysis by Dual Comparison With Cuff-Based Devices. *IEEE Access*, 13, 30587–30596.

[7] Martín-Escudero, P., Cabanas, A. M., Dotor-Castilla, M. L., et al. (2023). Are Activity Wrist-Worn Devices Accurate for Determining Heart Rate during Intense Exercise? *Bioengineering*, 10(2), 254.

[8] Nahar, J., Imam, T., Tickle, K. S., & Chen, Y. P. P. (2013). Computational intelligence for heart disease diagnosis: A medical knowledge driven approach. *Expert Systems with Applications*, 40(1), 96–104.

[9] Shrestha, D. (2024). Advanced Machine Learning Techniques for Predicting Heart Disease: A Comparative Analysis Using the Cleveland Heart Disease Dataset. *Applied Medical Informatics*, 46(3), 91–102.

[10] Thiyagaraj, M., & Suseendran, G. (2020). Enhanced prediction of heart disease using particle swarm optimization and rough sets with transductive support vector machines classifier. *Springer Nature*, 141–151. https://doi.org/10.1007/978-981-13-9364-8_11

Appendix A: R Code for Logistic Regression Analysis

The following R code was used to load and preprocess the dataset, fit the Reduced and Full logistic regression models, conduct diagnostic checks, evaluate model performance, and generate all figures presented in this report.

```
# ==============================================================================
# Heart Disease Analysis - Logistic Regression Framework
# ==============================================================================

# 1. SETUP & LIBRARY LOADING
# ------------------------------------------------------------------------------
if (!require("pacman")) install.packages("pacman")
pacman::p_load(tidyverse, car, broom, pROC, caret, splines)

# 2. DATA LOADING & PREPROCESSING
# ------------------------------------------------------------------------------
url <- "https://archive.ics.uci.edu/ml/machine-
learning-databases/heart-disease/processed.cleveland.data"

col_names <- c("age", "sex", "cp", "trestbps", "chol", "fbs", "restecg",
               "thalach", "exang", "oldpeak", "slope", "ca", "thal", "num")

heart_data <- read.csv(url, header = FALSE, col.names = col_names,
                       na.strings = "?")

clean_data <- heart_data %>%
  drop_na() %>%
  mutate(
    target = ifelse(num > 0, 1, 0),
    sex     = factor(sex, labels = c("Female", "Male")),
    cp      = as.factor(cp),
    fbs     = as.factor(fbs),
    restecg = as.factor(restecg),
    exang   = as.factor(exang),
    slope   = as.factor(slope),
    thal    = as.factor(thal),
    ca      = as.numeric(ca)
  )

# 3. MODEL SPECIFICATION
# ------------------------------------------------------------------------------
model_reduced <- glm(
  target ˜ age + sex + bs(thalach, df = 3) + trestbps + cp + exang,
  data = clean_data, family = binomial(link = "logit")
)
```

```
model_full <- glm(
  target ~ age + sex + bs(thalach, df = 3) + trestbps + cp + exang +
           restecg + oldpeak + slope + chol + fbs + ca + thal,
  data = clean_data, family = binomial(link = "logit")
)

calc_r2 <- function(model) {
  1 - (model$deviance / model$null.deviance)
}

print(calc_r2(model_reduced))
print(calc_r2(model_full))

# 4. STATISTICAL INFERENCE
# --------------------------------------------------------------------------------
anova(model_reduced, model_full, test = "Chisq")

model_stats <- bind_rows(
  glance(model_reduced) %>% mutate(Model = "Reduced (Consumer)"),
  glance(model_full)    %>% mutate(Model = "Full (Clinical)")
) %>% select(Model, AIC, BIC, deviance)

print(model_stats)

# 5. DIAGNOSTICS
# --------------------------------------------------------------------------------
print(vif(model_full))

# Cook's Distance
par(mfrow = c(1,1))
plot(model_reduced, which = 4, sub = "")
abline(h = 4/nrow(clean_data), col = "red", lty = 2)

# Linearity of the Logit
probabilities <- predict(model_full, type = "response")
probabilities <- pmin(pmax(probabilities, 1e-4), 0.9999)
logits <- log(probabilities / (1 - probabilities))

linearity_data <- clean_data %>%
  select(age, thalach, trestbps, oldpeak) %>%
  mutate(logit = logits) %>%
  pivot_longer(cols = -logit, names_to = "predictors",
               values_to = "predictor_value")

linearity_plot <- ggplot(linearity_data,
                         aes(x = predictor_value, y = logit)) +
  geom_point(alpha = 0.3) +
  geom_smooth(method = "loess", se = FALSE, color = "blue") +
  facet_wrap(˜ predictors, scales = "free_x") +
  theme_minimal() +
  labs(
    title = "Linearity Assumption Check: Continuous Predictors vs Logit",
    x = "Predictor Value",
    y = "Log-Odds (Logit)"
  )

print(linearity_plot)

# 6. PERFORMANCE METRICS (ROC & CONFUSION MATRIX)
# --------------------------------------------------------------------------------
prob_reduced <- predict(model_reduced, type = "response")
prob_full <- predict(model_full, type = "response")

roc_reduced <- roc(clean_data$target, prob_reduced, quiet = TRUE)
roc_full    <- roc(clean_data$target, prob_full, quiet = TRUE)

plot(roc_full, col = "blue",
     main = "ROC Comparison: Clinical vs Consumer Models")
plot(roc_reduced, col = "red", add = TRUE)

legend("bottomright",
       legend = c(
         paste("Clinical AUC:", round(auc(roc_full), 3)),
         paste("Consumer AUC:", round(auc(roc_reduced), 3))
       ),
       col = c("blue", "red"), lty = 1)

# Confusion Matrix (Threshold = 0.5)
pred_class <- ifelse(prob_reduced > 0.5, 1, 0)
conf_matrix <- confusionMatrix(factor(pred_class),
                               factor(clean_data$target),
                               positive = "1")

print(conf_matrix$byClass[c("Sensitivity", "Specificity",
                            "Precision", "Recall")])
print(conf_matrix$overall["Accuracy"])

# 7. MODEL INTERPRETATION
# --------------------------------------------------------------------------------
tidy(model_reduced, exponentiate = TRUE, conf.int = TRUE) %>%
  filter(p.value < 0.05) %>%
  select(term, estimate, p.value, conf.low, conf.high) %>%
  print()
```