Predicting Cardiovascular Risk Using the Framingham Heart Study: A Logistic Regression Model

Markus Baltzer Nielsen & Jahfar Houssein

2025-09-25

Exploratory data analysis

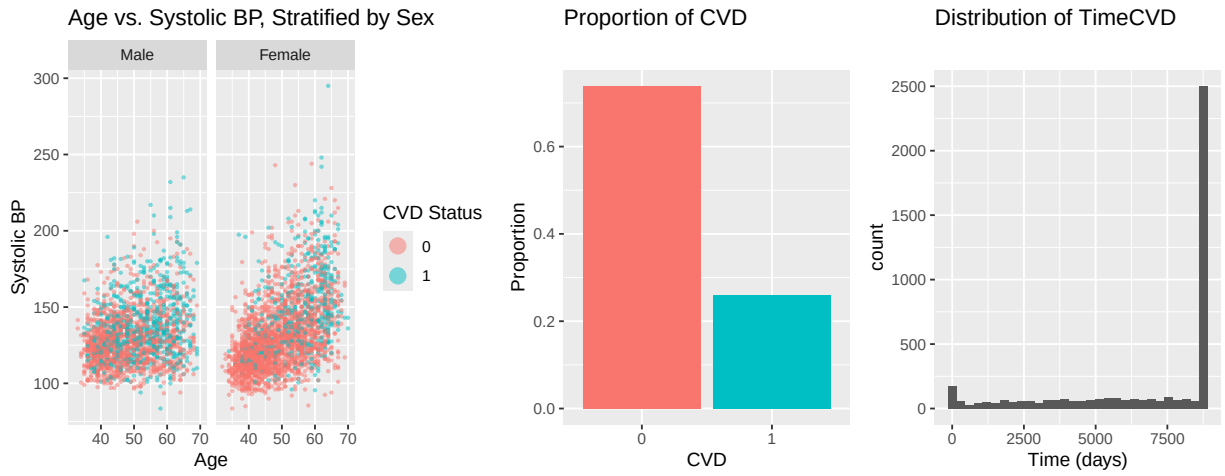
The Framingham Heart Study dataset contains 11,627 observations across 39 variables, representing a long-term cohort study of 4,434 individuals who were followed over a 24-year period, beginning in 1948. Participants underwent repeated examinations where a range of measures were recorded. We first load the framingham heart data and ensure the different columns are the right type of variable. To ensure clarity and clinical relevance in prediction, predictors X should be a fixed point in time. The baseline examination (PERIOD == 1) is the most logical choice, as it provides a consistent and interpretable starting point for all participants, with complete inclusion unlike later periods where some participants lack follow-up. This reduces the dataset to one row per individual, totaling 4434 observations. Consequently, given our focus on CVD (Cardiovascular disease during follow-up, 1 = occurred, 0 = did not occur) as the primary response variable in the binary regression and on TIMECVD(Time to CVD) in the survival model, we do not further investigate variables such as DEATH, ANGINA, STROKE, HYPERTEN, or other outcomes observed only at follow-up, nor their corresponding TIME variables. These variables serve as alternative response outcomes rather than predictors, and including them alongside CVD would introduce data leakage by incorporating information unavailable at the time of prediction for new patients with only baseline data. We acknowledge that the predictor PREVHYP(Prevalent Hypertension) is somewhat dependent on the second exam, as it reflects whether participants were treated for hypertension at baseline or had high blood pressure at the second exam, but we accept this limitation. Further description and analysis of the predictors will be conducted during the following exploratory data analysis phase.

These decisions establish a clear and consistent research question: Which baseline characteristics at the initial exam predict the risk of developing cardiovascular disease over the subsequent 24 years?

```
data(framingham, package = "riskCommunicator")
hjerte.data <- framingham |>
  filter(PERIOD == 1) |>
  dplyr::select(-HDLC, -LDLC) |> #HDLC and LDLC are for period=3
  mutate( CVD=as.numeric(CVD),
    SEX = factor(SEX, levels = c(1, 2), labels = c("Male", "Female")),
    educ = factor(educ,levels=c(1,2,3,4), labels=c("1","2","3","4")),
    across(
      .cols = c(BPMEDS, PREVCHD, PREVAP, PREVMI, PREVSTRK,
                PREVHYP,CURSMOKE,DIABETES),
      .fns = ~ factor(., levels = c(0, 1))
    )
)
```

We see that binary variables are coded as 1 if "yes" to the disease or condition and 0 if "no". Let's first examine the marginal distribution of our response variable to gain an initial understanding of its characteristics and how it relates to predictors such as age and systolic blood pressure (SYSBP).

```
prop<-hjerte.data |>
  count(CVD) |>
  mutate(proportion = n / sum(n))
```



The scatterplot demonstrates that age, systolic blood pressure, and male sex are seemingly positively associated with increased CVD risk. Participant RANDID = 1080920 exhibits an exceptionally elevated SYSBP, consistent with other clinical indicators, including a markedly high BMI (body mass index). Similarly, RANDID = 6300384 shows the highest BMI (56.8), with other vital signs in concordance, suggesting the absence of measurement error. Internal data validation checks were conducted (e.g., CIGPDAY(Cigarettes per day) > 0 only when CURSMOKE(current smoker) $= 1$), all of which were satisfied. The dataset exhibits substantial class imbalance, with the majority of participants, approximately 74%, remaining free of CVD event, as reflected in the observed proportions. Time of CVD cluster near the study's endpoint, indicating a high proportion of right-censored observations. Next, we will examine the missing values.

```
na_percent <- colSums(is.na(hjerte.data)) / nrow(hjerte.data) * 100
na_percent <- na_percent[na_percent > 0]
```

Table 1: Percentage of Missing Data per Variable

TOTCHOL	CIGPDAY	BMI	BPMEDS	HEARTRTE	GLUCOSE	educ
1.17%	0.72%	0.43%	1.38%	0.02%	8.95%	2.55%

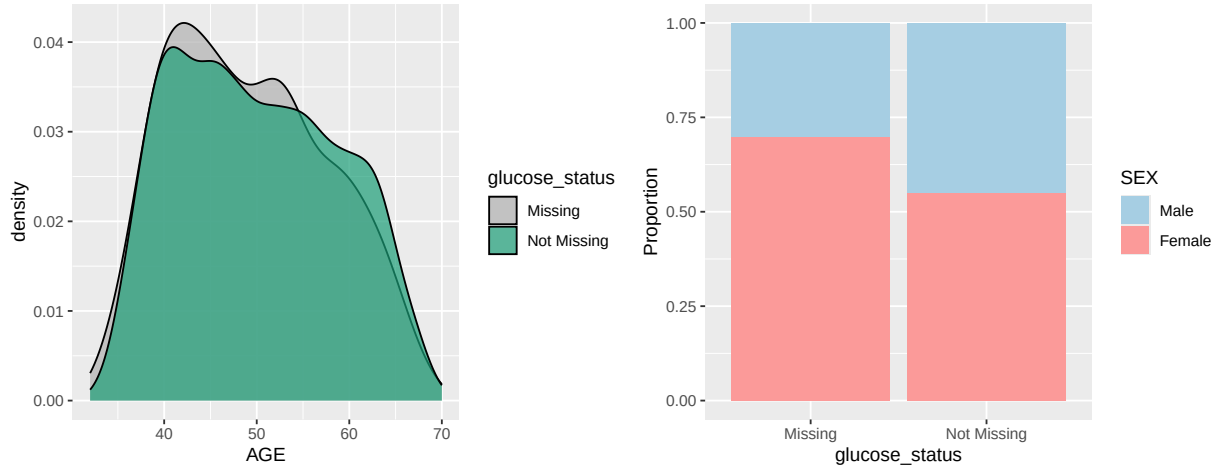
Missing data are present across several variables, though most columns exhibit less than 2–3% missingness and are considered negligible after further analysis. Notably, CIGPDAY is missing exclusively for smokers, who display higher CVD rates, however, the small sample limits conclusions. BPmeds (Blood pressure medication) shows slightly higher CVD proportions among missing values, but differences are not statistically significant. Subsets of other variables with missing values exhibit the same CVD prevalence as those without missing data, therefore possibly MAR, and thus unlikely to bias predictors upon removal. The primary concern is GLUCOSE(Casual serum glucose), with ~9% missing values, representing the most prevalent missing predictor. Furthermore, TOTCHOL(Total Cholesterol) is missing in 80% of cases with missing Glucose. To assess potential impact, CVD prevalence is compared between observations with missing versus complete Glucose data.

```r
analysis_data <- hjerte.data |>
mutate(glucose_status = ifelse(is.na(GLUCOSE), "Missing", "Not Missing"))
```

Table 2: Proportions of CVD by glucose status

	Not missing glucose	Missing glucose
rows	4037.00	397.0
count_cvd	1077.00	80.0
proportion_cvd	0.27	0.2

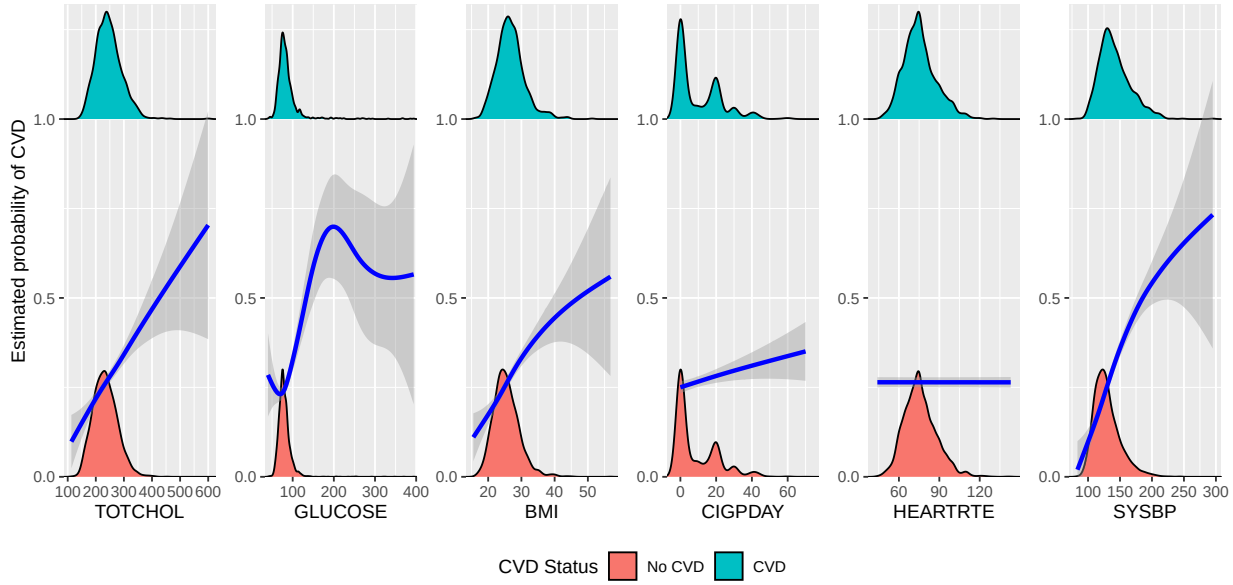
The discrepancy between missing CVD proportion (20.2%) and non-missing (26.7%) data suggests that excluding NA values could overestimate CVD prevalence.



Upon further examination, the distribution of the glucose-missing subset appears largely consistent with the non-missing data across most predictors, except for age and sex, as illustrated in the plot above, with missingness being predominantly among younger females. This observation is noteworthy, as the initial scatterplot suggested that higher age and male sex were associated with increased CVD incidence. Consequently, the missingness in glucose could bias the overall CVD prevalence if it is influenced by these variables. However, if age and sex fully account for the missingness mechanism, it can be classified as Missing at Random (MAR), in which case excluding the missing values may be reasonably justified without substantial bias.

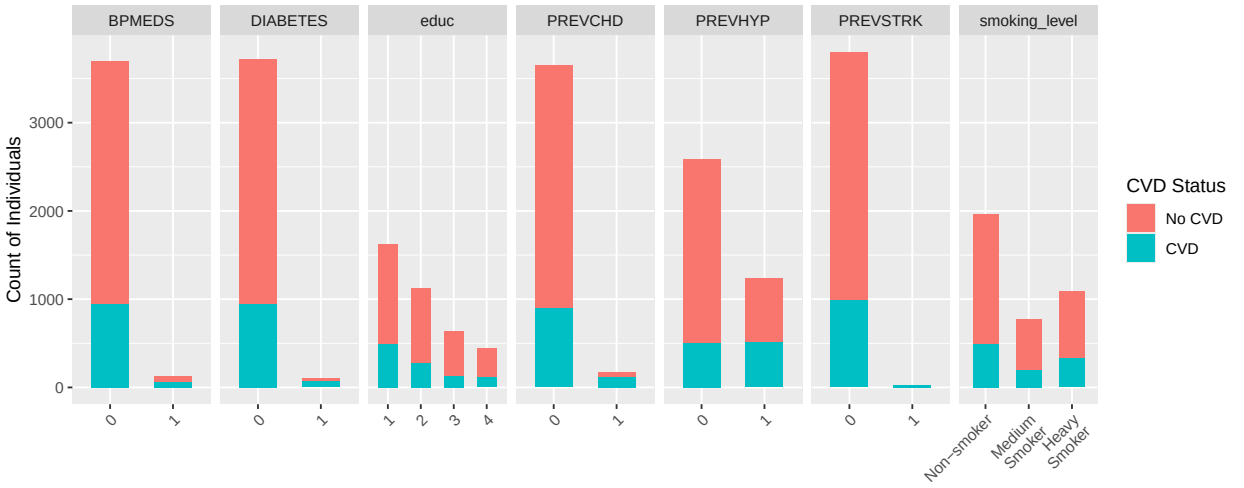
After stratifying by sex and age to ensure comparable distributions, CVD proportions became more consistent across strata, reducing concerns that missing glucose values may be systematically related to an increased or decreased CVD risk. Missingness was more frequent among younger individuals, particularly females, suggesting a plausibly MAR mechanism driven by age and sex. Notably, in the full dataset, extreme glucose values (both high and low) are correlated with changes in other vital signs, a pattern not seen in the missing subset. This supports the idea that the missing data are not missing due to glucose extremes (i.e., not MNAR). Following the thorough examination of the missing predictor variables and their potential impact, we will proceed with the complete observed dataset. As imputation is out of the scope of the course, although briefly mentioned, we refrain from doing this. Although this decision results in roughly a 15% reduction in sample size, The exclusion primarily affects younger women whose risk profiles are comparable to those of individuals with non-missing observations. Thus, while the baseline may reflect a slightly higher baseline CVD prevalence due to the reduced representation of low-risk individuals, the parameter estimates remain unbiased as the excluded group mirrors the retained cohort in key covariates. Let's proceed to examine the predictors using the complete dataset.

```
hjerte<-na.omit(hjerte.data)
```

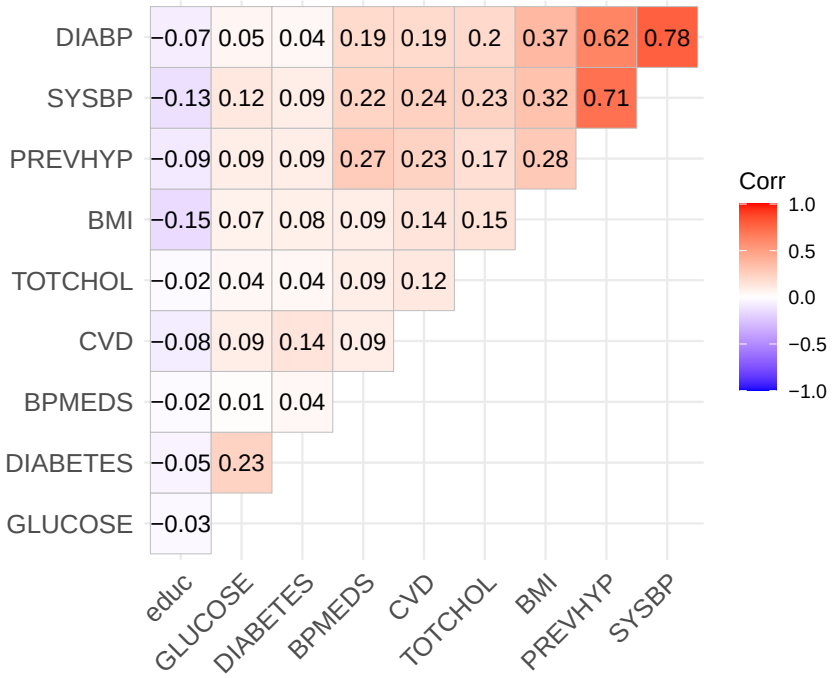


We visualize each predictor using density plots stratified by CVD status to compare distributions across outcome levels. To assess associations with CVD, we overlay a GAM-based geom_smooth(), estimating the conditional mean of CVD as a function of each predictor. This flexible approach captures potential non-linear relationships, making fewer parametric assumptions at this EDA stage to explore associations. While a logistic GLM could constrain fitted values to [0,1], the GAM allows this exploratory flexibility, with minor excursions beyond [0,1] for confidence intervals for SYSBP and TOTCHOL. For most continuous predictors, higher values correspond to increased CVD probability, except HEARTRTE, with greater uncertainty in data sparse tails. The smoothed conditional mean suggests glucose, and potentially BMI and SYSBP, may benefit from spline modeling. CIGPDAY, being zero-inflated and right-skewed, is collapsed into quantile-based categories (non-smoker, moderate, heavy) to better reflect smoking intensity than the binary CURSMOKE variable.

```
hjerte <- hjerte %>%
  mutate(smoking_level = factor(case_when(
    CIGPDAY == 0 ~ "Non-smoker", CIGPDAY < 20 ~ "Medium Smoker",
    CIGPDAY >= 20 ~ "Heavy Smoker"
  ), levels = c("Non-smoker","Medium Smoker","Heavy Smoker")))
```



Educ(education), like female sex as previously hypothesized, appears to serve as a protective factor—particularly at higher attainment levels. Prevalent conditions such as coronary heart disease (CHD), PREVSTRK(stroke), PREVHYP(hypertensive), diabetes, and heavy smoking seem to be risk factors for CVD. For example, prevalent stroke "seperates" CVD in this dataset, as all individuals with stroke history also experienced a CVD event. However, if included as predictor in a model, it leads to unstable parameter estimates due to convergence issues of the MLE. To improve interpretability and ensure adequate sample size, we focus on prevCHD, an umbrella variable encompassing related heart conditions (PREVMI, PREVAP). These subcategories are conceptually similar but too sparse for stable estimation. PREVMI==1 also nearly perfectly predicts CVD (76/78 cases), possibly complicating estimation. While limited variability in other conditions may pose issues, sufficient case counts (e.g., 128 with diabetes) help mitigate this. The collapsed CIGPDAY variable provides adequate group sizes, improving both model stability and interpretability.



Having established associations between most predictors and CVD prevalence, we limit the correlation plot to key variables. Using a Spearman correlation matrix, we assess potential collinearity as a diagnostic for estimation and interpretability issues. The matrix, adapted from the Week 2 Cecilie solutions, confirms known CVD predictors but highlights notable multicollinearity. Notably, there are strong intercorrelations

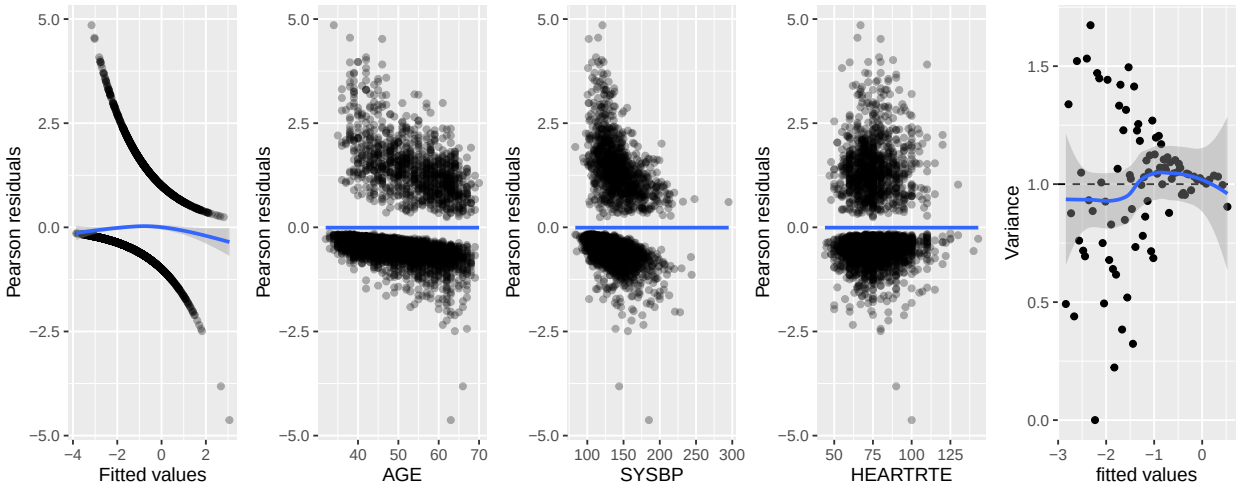
among the blood pressure–related measures (SYSBP, DIABP, PREVHYP), with BPMEDS also showing some correlation with these variables. To a lesser extent, GLUCOSE and DIABETES are similarly correlated. Such high intercorrelations indicate redundancy, which could bias estimates and obscure the individual effects of each predictor on CVD risk. To address this, we may remove redundant variables to reduce multicollinearity and ensure more reliable estimates of each predictor's contribution, which may improve model interpretability.

Binary Regression model

we are attempting to build a model capable of prediciting probability of Y=CVD=1, given health indicators $X = x$, that is, $p(x) = P(Y = 1|X = x)$. This approach aligns with the binary regression framework. We begin by fitting a full additive model using the binomial family, whose canonical link function is the logit (see Exercise 5.2, RwR). While we acknowledge that other link functions such as the probit or complementary log-log (cloglog) are available, they typically exert negligible influence on fitted probabilities (Chapter 10, RwR). The cloglog link, being asymmetric (figure 10.4 RwR), could be considered more appropriate given the 27% base prevalence of CVD. This asymmetry is useful because it does not assume that a predictor's effect is symmetrical around the 50% probability point. Instead, it can better model scenarios where risk accelerates, meaning an increase in a risk factor has a larger impact on individuals already at moderate risk than on those at very low risk. As shown in Figure 10.4 RwR, the cloglog link approaches probability 1 more rapidly and 0 more gradually, making it potentially preferable for imbalanced data with rare CVD cases. However, we value the interpretative clarity of the logit link in this context. Logistic regression coefficients represent log-odds ratios, which can be exponentiated to yield odds ratios, providing a clear and familiar measure of how risk factors such as age or smoking affect disease risk. The logit link also ensures predicted probabilities remain within the (0, 1) range. The variable PREVSTRK is excluded to avoid convergence issues in the Fisher scoring algorithm, as previously noted and PREVCHD includes PREVMI and PREVAP. Our objective is to develop a model that accurately predicts the occurrence of cardiovascular disease (CVD) within a 24-year follow-up period based on current health indicators and known comorbidities.

```
form<-CVD ~AGE+SEX+TOTCHOL+BMI+smoking_level+
    DIABETES+PREVHYP+PREVCHD+HEARTRTE+BPMEDS+SYSBP+DIABP+GLUCOSE+educ
basemodel<- glm(form,family=binomial, data=hjerte)
small_glm_aug <- augment(basemodel, type.residuals = "pearson",data=hjerte)
```

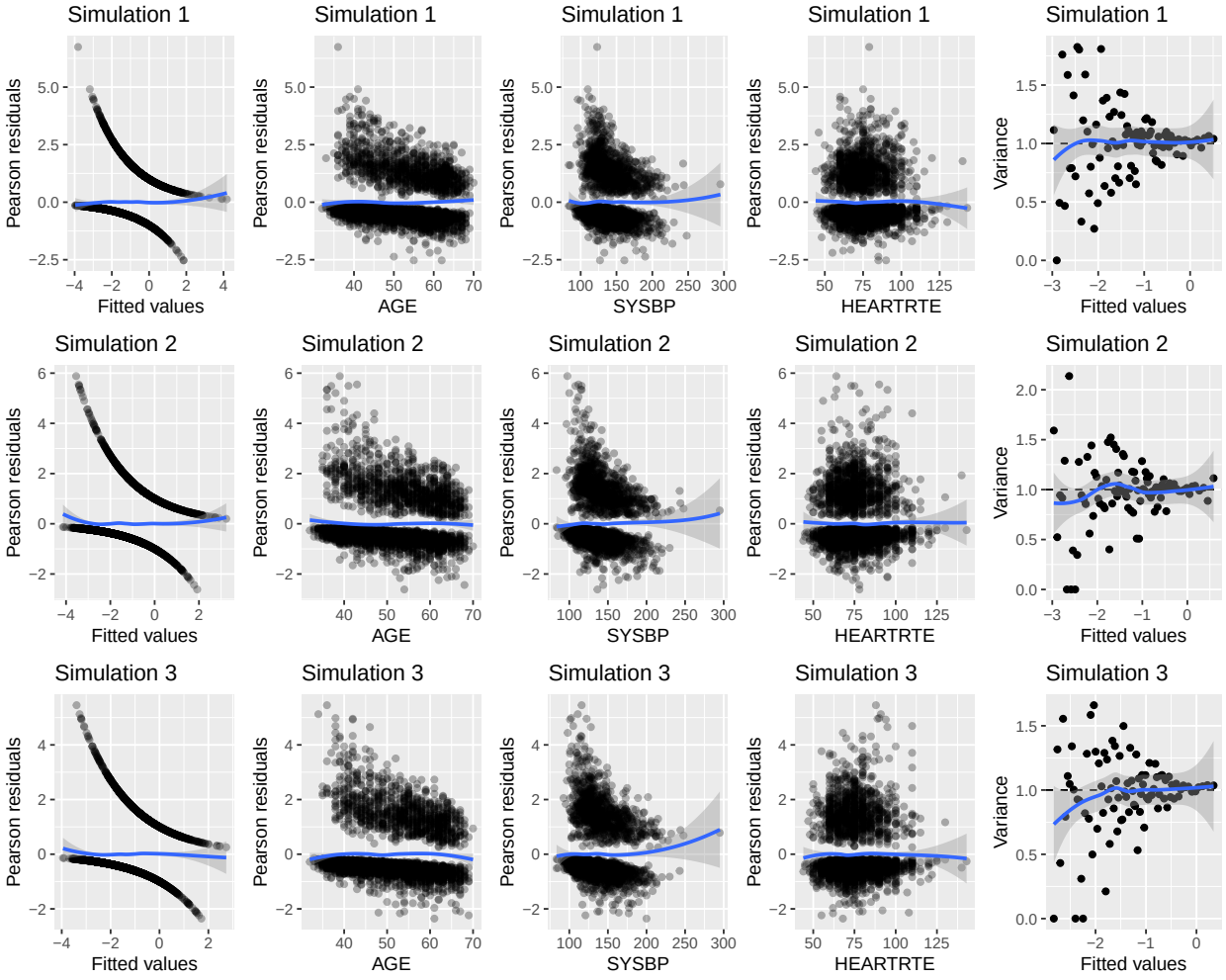
Lets perform model diagnostics on the above firstly, using code from chapter 10 in RwR.



The smoothed line in the residuals-versus-fitted and residuals-versus-predictors plots remains close to zero, exhibiting no evident nonlinear patterns, indicating that the model's mean—and therefore variance—structures are appropriately specified (Chapter 10, RwR). To further assess variance, we plot the locally

estimated Pearson residual variances against the fitted values, using the same code and binning strategy as in Chapter 10. The fitted values are on the link scale, specifically the log-odds scale. These local variances provides a supplementary check that the mean model is correctly specified, though it is not an independent test of the variance function itself. The funnel-shaped patterns for AGE and SYSBP arise naturally from the logistic model and their positive correlations with CVD: as these predictors increase, the predicted probability of CVD rises, systematically altering residual magnitudes and creating the curved pattern. The funnel's asymmetry—a dense lower band of negative residuals and a sparse upper curve of positive ones—is a direct reflection of class imbalance. The many CVD-free participants form a dense cluster of small negative residuals, while the few who developed CVD despite low predictor values create a sparse set of large positive residuals. As predictor values rise, these positive residuals decrease because the model's predicted probability of CVD increases, reducing residual magnitude for CVD = 1 but increasing it for CVD = 0. This funnel also appears in the local variance plot for the same reason, amplified by class imbalance. On the left, bins with low predicted log odds mix frequent small negative residuals (CVD=0) with rare large positive ones (CVD=1), producing unstable, scattered variance estimates. On the right, the band becomes tighter, showing that as predicted log-odds increase, residual variance decreases. This occurs because there are fewer false negatives and more correctly predicted positives (CVD = 1), leading to more stable model estimates.

This interpretation is consistent with the HEARTRATE residuals, which do not exhibit a funnel pattern. This absence arises from the weak association, as seen previously, assigning nearly uniform probabilities across its domain. However, the residuals still display noticeable asymmetry, a consequence of class imbalance. To validate the residuals further, we will simulate from the fitted model and compare residuals to the observed results. We use Cecilie's simulation code from Week 4.



Simulated residuals are largely consistent, though the mean function of SYSBP may be influenced by the large, physiologically plausible outlier, as discussed previously. Nonlinear modeling approaches, such as splines, will be explored to better capture SYSBP structure. While some degree of uncertainty always remains, the diagnostic plots indicate that the logistic regression model adequately captures the underlying data structure. For now, predictor significance will be assessed via the deviance test.

```
drop1(basemodel, test = "LRT") |>
  tidy() |>
  arrange(p.value, desc(LRT)) |>
  filter(term != "<none>") |>
  dplyr::select(-AIC) |>
  knitr::kable() |>
  kable_styling(full_width =FALSE,font_size=10)
```

term	df	deviance	LRT	p.value
SEX	1	3849.531	91.9251204	0.0000000
AGE	1	3829.407	71.8012363	0.0000000
PREVCHD	1	3817.286	59.6795835	0.0000000
TOTCHOL	1	3781.643	24.0364421	0.0000009
smoking_level	2	3778.752	21.1458164	0.0000256
HEARTRTE	1	3769.855	12.2487558	0.0004656
DIABETES	1	3769.129	11.5230752	0.0006874
PREVHYP	1	3767.614	10.0074430	0.0015591
BMI	1	3765.709	8.1025353	0.0044203
SYSBP	1	3763.377	5.7704668	0.0162977
GLUCOSE	1	3761.057	3.4508362	0.0632198
educ	3	3764.739	7.1326527	0.0677877
DIABP	1	3760.412	2.8056383	0.0939335
BPMEDS	1	3757.643	0.0364338	0.8486224

The non-significance of DIABP and BPMEDS likely reflects collinearity with other blood pressure measures, as previously reported. Consequently, the hypothesis supporting their inclusion is rejected. The possible exclusion of DIABP and BPMEDS is further supported by prior evidence of multicollinearity with SYSBP (Regression, 2020, p. 94). We therefore remove them for now in a reduced model. Glucose and education exhibit borderline significance, glucose likely due to its association with diabetes, but are retained for now. Variables are reordered by statistical significance, and parameter estimates with confidence intervals are examined.

```
form_reduced<-CVD ~ SEX + AGE + PREVCHD + TOTCHOL + smoking_level + HEARTRTE+
DIABETES + PREVHYP + BMI + SYSBP + GLUCOSE + educ
basemodel_reduced<- glm(form_reduced,family=binomial, data=hjerte)
tidy(basemodel_reduced)[-1, "estimate"] |>
    cbind(confint.lm(basemodel_reduced)[-1, ]) |>
    knitr::kable(digits = 3)
```

	estimate	2.5 %	97.5 %
SEXFemale	-0.871	-1.046	-0.696
AGE	0.045	0.034	0.055
PREVCHD1	1.363	1.003	1.722

	estimate	2.5 %	97.5 %
TOTCHOL	0.005	0.003	0.006
smoking_levelMedium Smoker	0.359	0.141	0.576
smoking_levelHeavy Smoker	0.417	0.218	0.617
HEARTRTE	-0.012	-0.019	-0.005
DIABETES1	0.921	0.379	1.463
PREVHYP1	0.397	0.175	0.618
BMI	0.034	0.013	0.054
SYSBP	0.011	0.006	0.016
GLUCOSE	0.004	0.000	0.008
educ2	0.121	-0.075	0.316
educ3	-0.195	-0.440	0.051
educ4	-0.122	-0.387	0.144

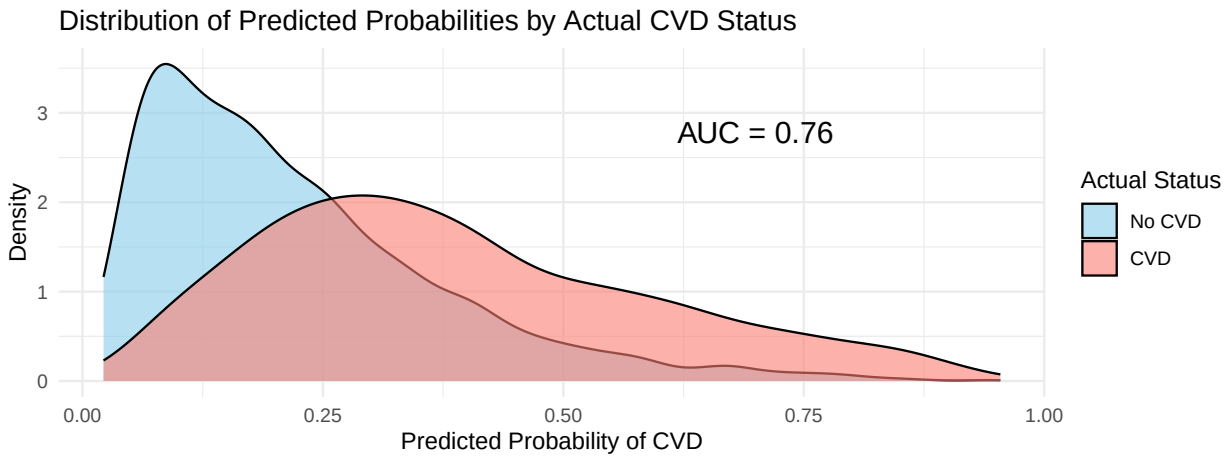
Examination of the confidence intervals from the fitted model identifies several significant predictors of cardiovascular disease (CVD), as the majority of intervals exclude zero. The strongest risk factors are a history of coronary heart disease (PREVCHD) and the presence of diabetes. Quantitatively, a history of CHD is associated with an odds ratio (OR) of approximately $\exp(1.363) = 3.91$, meaning an individual with pre-existing CHD has nearly four times the odds of developing CVD. Using the overall sample prevalence of 26% as an illustrative baseline, this would escalate an individual's risk to approximately 58%:

$$\hat{p} = \frac{OR \cdot p}{1 - p + (OR \cdot p)} = \frac{3.91 \cdot 0.26}{1 - 0.26 + (3.91 \cdot 0.26)} \approx 0.58$$

It is important to note that using the overall sample prevalence as the baseline probability is a simplification. The true baseline risk for the reference group,an individual without pre-existing CHD,would inherently be lower than 26%. Thus, this calculation primarily serves to demonstrate the large multiplicative effect of the odds ratio. Conversely, female sex is the most significant protective factor, reducing the odds of CVD by approximately 58% (OR 0.42, 95% CI 0.35-0.50) compared to males. The model shows a dose–response relationship for smoking, with higher odds of CVD among medium (OR 1.43, 95% CI 1.15–1.78) and heavy smokers (OR 1.52, 95% CI 1.24–1.85) compared to non-smokers. The overlapping confidence intervals indicate that the difference between these groups may not be statistically significant, though the overall trend suggests increasing risk with smoking intensity. Additional predictors such as increasing age, systolic blood pressure, and BMI also contribute to elevated risk. The effect of education, however, remains tentative, as its confidence interval includes zero, aligning with its non-significance in the deviance test.

We will now turn our heads to evaluate the models predictive performance. We use chapter 10 code to define a function to calculate AUC.

```
model_data <- augment(basemodel_reduced, type.predict = "response")
auc <- function(y, eta) {eta1 <- eta[y == 1];eta0 <- eta[y == 0]
    wilcox.test(eta1, eta0)$statistic / (length(eta1) * length(eta0))}
```



The density plot shows the model's discriminative ability. It performs well for the "No CVD" group, assigning low predicted probabilities, but shows mixed results for the "CVD" group, with substantial overlap between 0.10 and 0.50 probabilities, reflecting uncertainty in distinguishing high-risk from low-risk individuals. This moderate separation is reflected in an AUC of 0.76, indicating fair but improvable performance.

Model Enhancements

We have already developed a reduced model; however, we will explore the addition of further complexity by investigating interaction effects and incorporating spline functions. To avoid unstable coefficient estimates, DIABP and BPMEDS were also excluded from consideration in models featuring splines and interactions. Based on these insights, we specified three competing Generalized Linear Models (GLMs) with a binomial family and logit link function to estimate the risk.

- The Additive Model (basemodel_reduced): Our baseline model with only additive terms however we have removed the 2 variables.
- The Full Pairwise Interaction Model (model_int): This complex model was constructed to explore potential effect modification, where the effect of one predictor may depend on the level of another. However, its large number of parameters carries a significant risk of overfitting the training data this will be explored.
- The Natural Spline Model (model_spline): This model specifically addresses the suspected non-linearities observed earlier in EDA with the smoothed conditional means. by fitting natural splines with 4 degrees of freedom to GLUCOSE, BMI and SYSBP, while treating all other predictors additively. This allows the model to capture the slight curve in predictors observed earlier in our analysis.

The new model formulas were defined in R as follows:

```
form_int <- CVD ~ (AGE+SEX+TOTCHOL+BMI+smoking_level+PREVHYP+PREVCHD+
  HEARTRTE+GLUCOSE+SYSBP+DIABETES+educ)^2

model_int <- glm(form_int, family = "binomial", data = hjerte)

form_spline <- CVD ~ ns(BMI, df=4) + ns(SYSBP, df=4)+ns(GLUCOSE, df=4) + SEX + TOTCHOL + AGE +
smoking_level + DIABETES + PREVHYP + PREVCHD + HEARTRTE  + educ

model_spline <- glm(form_spline, family = "binomial", data = hjerte)
```

Goodness of fit

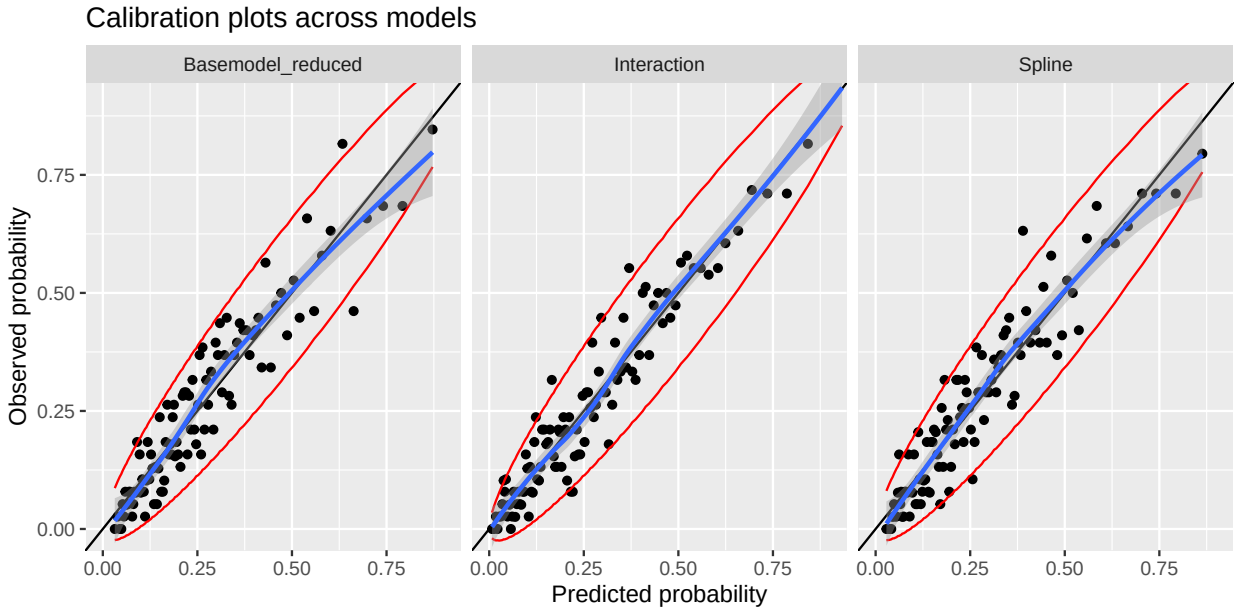
To select the most parsimonious and well-fitting model, we use the Akaike Information Criterion (AIC), defined as $AIC = -2\log(\text{likelihood}) + 2k$, where k is the number of parameters. The AIC penalizes model complexity and is supported by two error measures: the Deviance loss ($\text{err} = \frac{D}{n}$) and the Pearson loss ($\text{err} = \frac{\chi^2}{n}$), which assess the adequacy of the binomial error structure and potential overdispersion. The comparative results for the three models are presented in the table below.

Table 5: AIC, Pearson and Deviance

Model	AIC	Pearson	Deviance
Basemodel_reduced	3792.44	0.97	0.98
Interaction	3832.47	0.99	0.94
Spline	3789.64	0.98	0.98

The dispersion diagnostics confirm that the model structure is appropriate. The Pearson and Deviance losses, representing mean squared residuals, are tightly clustered near **1.0** (ranging from 0.94 to 0.99), indicating no overdispersion and validating the binomial error assumption. Consequently, AIC serves as a reliable criterion for model comparison. The Interaction model is clearly penalized for its complexity, with an AIC of **3832.47**—substantially higher than the simpler models—suggesting overfitting without improved explanatory power. The Spline model yields the lowest AIC (**3789.64**), only marginally better than the Basemodel_reduced (**3792.44**), a difference of just **2.8**. Given this small margin, the Spline model represents the best balance between parsimony and predictive accuracy, justifying inclusion of the non-linear term.

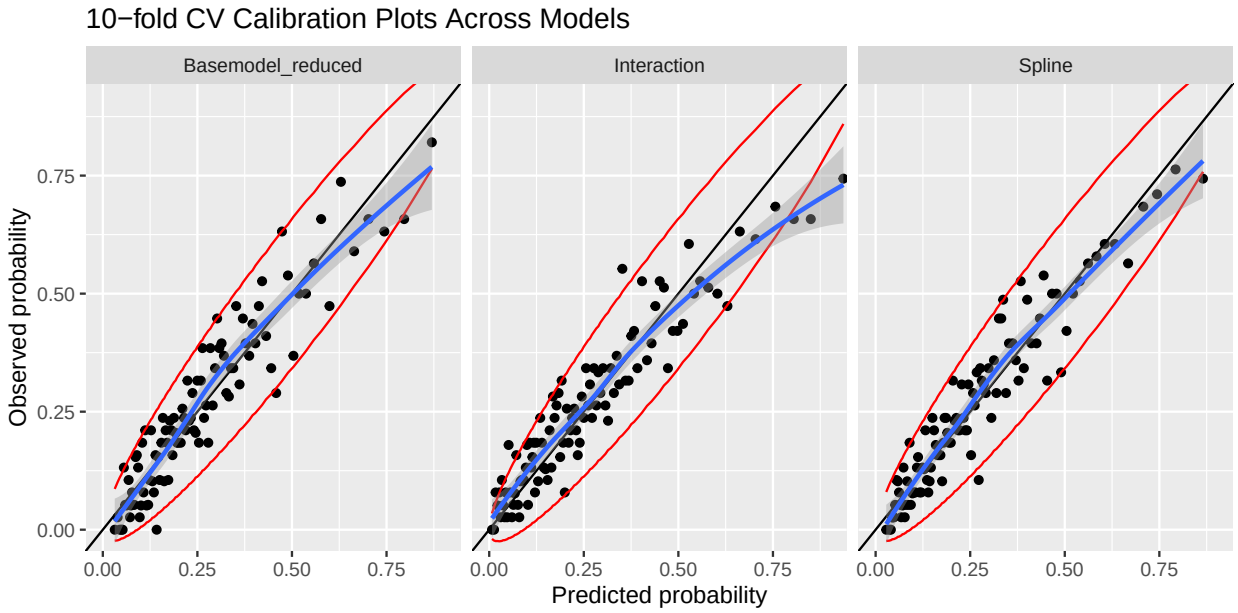
We will take a look at the calibration graphs to see how well the probabilities are aligned to the observed frequencies. The code is a copy of the code found in the coursebook chapter 10.



The figure above are calibration graphs based on the training data, we see that all models appear exceptionally well-calibrated. The points cluster tightly around the ideal y = x diagonal line, and the smoothed curves track it almost perfectly. However, this apparent accuracy can be an artifact of overfitting, especially for the flexible Interaction model.

We will perform 10-fold cross-validation by defining a (cv) function and then using this to generate cross-validated calibration plots, like in chapter 10 rwr.

```
cv <- function(form, data, K = 10) {
  n <- nrow(data)
  response <- all.vars(form)[1]
  mu_hat <- numeric(n)
  group <- sample(rep(1:K, length.out = n))
  for(i in 1:K) {
    model <- glm(form, data = data[group != i, ], family = "binomial")
    mu_hat[group == i] <-
      predict(model, newdata = data[group == i, ], type = "response")
  }
  mu_hat
}
```



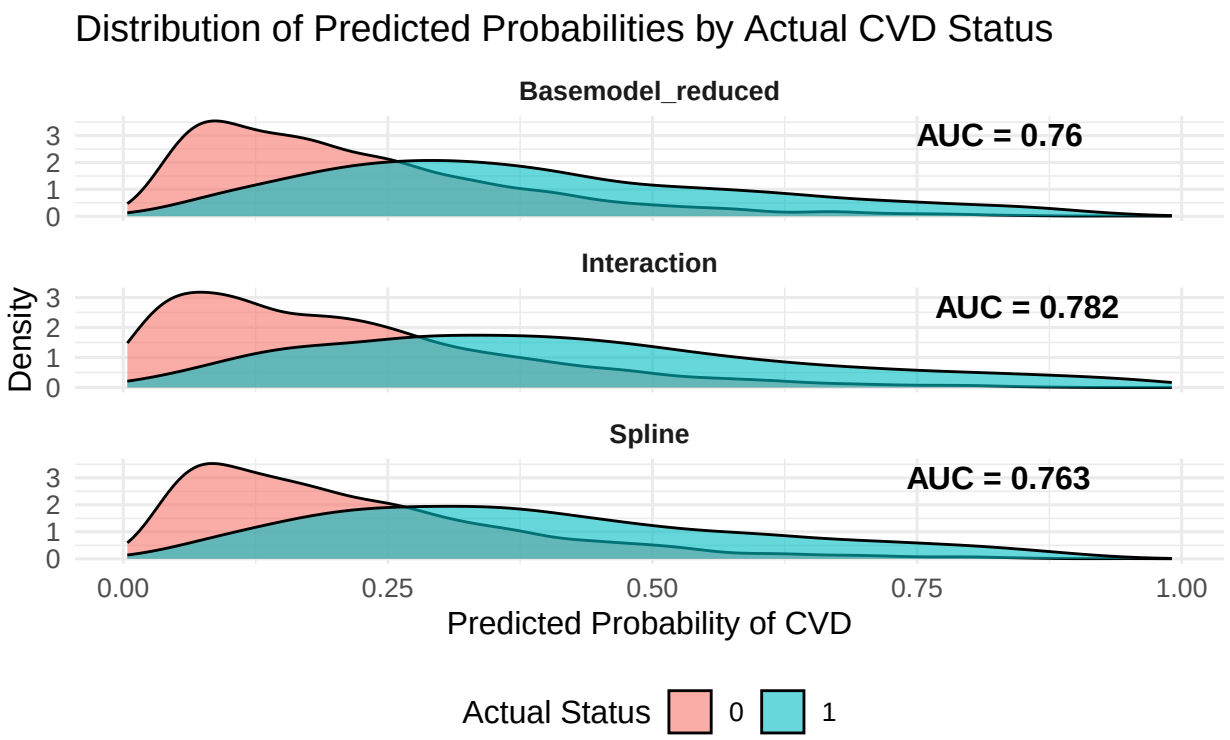
The 10-fold cross-validated calibration plots above, provide a more honest assessment of performance on unseen data. Here, systematic miscalibration becomes apparent:

- Overestimation at High Risk: All three models consistently overestimate risk for individuals with a predicted probability greater than 0.5. This is visible where the blue smoothed curve falls below the diagonal, indicating that the observed event rate is lower than predicted.
- Underestimation at Moderate Risk: Conversely, there is a slight underestimation for probabilities in the 0.2 to 0.4 range.

The confidence bands (grey region) are a little wider in the CV plots, reflecting higher uncertainty. Critically, the Interaction model's calibration curve deviates more significantly and shows greater instability than the Basemodel_reduced and Spline models. This assessment reinforces the AIC results: the simpler models, being less prone to overfitting, provide more reliable and generalizable probability estimates.

AUC and cross-validation

We will now take a look at the predictive performance by looking at how well the model discriminates between the CVD levels.



Above is a graph of the 3 models with the AUC values attached to the top right this has been calculated by using the AUC function from chapter 10 in rwr books. Discrimination is assessed via AUC (concordance measure, robust to imbalance), with density plots of predicted probabilities by CVD status. Initial AUCs: **0.76** (Basemodel_reduced), **0.782** (Interaction), **0.763** (Spline). The interaction's higher AUC reflects better separation (green curve shifted right therefor less overlap), but the cross-validated AUC is quite different as seen below:

Table 6: Cross-validated AUC

Model	AUC	Pearson	Deviance
Basemodel_reduced	0.755	0.992	0.993
Interaction	0.739	1.246	1.034
Spline	0.755	1.009	0.993
Basemodel	0.753	1.003	0.991

Again we have used the cv-function defined earlier including the pearson- and deviance -loss. It is apparent based on the interaction's AUC decline of **0.043** and increase in error loss when we compare with table 5, that the previous results were optimistic and doesn't generalize to unseen data well. Furthermore, the other 3 models become almost identical, therefor we choose the Basemodel_reduced, favoring parsimony.

Threshold analysis

The final step is to determine the optimal probability cut-off, or threshold (t), for the selected Base-model_reduced Model. We know from the coursebook chapter 9: Section 9.6, that while the overall discrimination (AUC) is model-invariant to the threshold, the selection of t directly impacts the trade-off between false negatives (missed cases) and false positives (unnecessary alarms). The dataset is imbalanced, with only **26**.**4**% of individuals experiencing a CVD event (73.6% CVD-free). This imbalance can bias a model's performance, particularly when using a default 0.5 classification threshold, and necessitates careful tuning and evaluation based on metrics sensitive to false negatives.

Given the clinical context of CVD prediction, where the objective is to identify high-risk patients for preventative intervention, the cost of a False Negative (FN)—failing to classify an at-risk person—is high. This drives a need to optimize metrics that favor case identification, such as Recall (Sensitivity).

We evaluated the model's performance metrics (Recall, Precision, and the F1-score) across a range of thresholds from $t = 0.10$ to $t = 0.50$ with jumps of 0.05 using our cross validation (cv) function from earlier.

```
set.seed(123456)
prob_base <- cv(form_reduced, hjerte, 10)
y <- hjerte$CVD
threshold <- seq(0.1,0.5,0.05)
recall <- precision <- f1_score <- numeric(length(threshold))

for(i in seq_along(threshold)){
  t <- threshold[i]; pr <- ifelse(prob_base>t,1,0)
  TP <- sum(pr==1 & y==1); FP <- sum(pr==1 & y==0); FN <- sum(pr==0 & y==1)
  recall[i] <- ifelse(TP+FN>0, TP/(TP+FN), NA)
  precision[i] <- ifelse(TP+FP>0, TP/(TP+FP), NA)
  f1_score[i]  <-
    ifelse(!is.na(recall[i])&!is.na(precision[i])&(recall[i]+precision[i])>0,
                  2*recall[i]*precision[i]/(recall[i]+precision[i]), NA)
}
```

Table 7: Cross-validated metrics across thresholds

Threshold	Recall	Precision	F1_Score
0.10	0.956	0.312	0.471
0.15	0.900	0.350	0.504
0.20	0.813	0.394	0.531
0.25	0.718	0.434	0.541
0.30	0.605	0.469	0.528
0.35	0.505	0.503	0.504
0.40	0.400	0.526	0.454
0.45	0.324	0.572	0.414
0.50	0.272	0.608	0.376

The F1-score, which is the harmonic mean of Precision and Recall, provides a single metric for optimizing a balanced performance between identifying positive cases (high Recall) and minimizing false alarms (high Precision).

The results in Table 7 show that the F1-score is maximized at a threshold of $\mathbf{t = 0.25}$ (F1-score = **0**.**541**).

- Trade-off Interpretation: At $t = 0.25$, the model achieves a high Recall of **71**.**8**%, meaning it successfully identifies over two-thirds of the actual CVD cases. This is a significant improvement over the default

$t = 0.50$ (where Recall is only **27.2**%), and directly aligns with the objective of reducing costly False Negatives. This gain in Recall comes at the expense of Precision, which stands at **43.4**%, indicating that slightly more than half of the patients flagged as "high risk" by the model will not actually develop CVD within the follow-up period. But the people that are false positive have a pretty high probability of developing CVD according to our model, and this could also be beneficial knowledge.

In summary, choosing a threshold of **0.25** provides the best overall balance between catching true disease cases and maintaining a clinically acceptable rate of false positives. Any predicted probability $\hat{\mathbf{p}} \geq \mathbf{0.25}$ is therefore classified as a positive CVD prediction.

Table 8: False Positives and False Negatives at Thresholds 0.25 and 0.5

Threshold	FP	FN
0.25	946	285
0.50	177	736

The above table shows how big a difference it made. it shows the count of false positives (FP) and the false negatives (FN), for both the old threshold of 0.5 and the new threshold of 0.25. We see that the model now serves our purpose better by decreasing the false negatives by quite a big margin, of cause this also increases the false positives and this is where we use the F1_score to balances this tradeoff.