2a

Let X be the cost of a person on medication, and $\bar{X}$ of a person without medication. Then taking the expected value gives

$$\begin{aligned} &E[X] = -172500p/2.3 - 7500(1 - p/2.3) = -71739p - 7500 \\ &E[\bar{X}] = -165000p \end{aligned} \tag{1}$$

From an economic point of view, you should prescribe medication when $E[X] > E[\bar{X}]$ in order to minimize the expected loss of the outcome

$$\begin{aligned} -71739p - 7500 &> -165000p \\ p &> 0.0804 \end{aligned} \tag{2}$$

So we should prescribe medication when p is over 8.04%.

2b

```r
model2b = glm(TenYearCHD ~ ., family="binomial", data=train)
```

From the summary of the model we can see that gender, age, cigarettes per day, total cholesterol, systolic blood pressure and blood glucose level is staistically significant. They all have positive coefficients, so the risk increases with these factors. Most of these, like gender, age, cigarettes per day, cholestetrol and blood pressure are "known" for increasing risk of heart diseases, so these results are not very surprising. I am, however, a bit surprised that BMI is not more significant.

c

```r
pasient1 = data.frame(
  male = "1",
  age = 55,
  education = "College",
  currentSmoker = "1",
  cigsPerDay = 10,
  BPMeds = "0",
  prevalentStroke = "0",
  prevalentHyp = "1",
  diabetes = "0",
  totChol = 220,
  sysBP = 140,
  diaBP = 100,
  BMI = 30,
  heartRate = 60,
  glucose = 80
)
risk = predict(model2b, newdata=pasient1, type="response")
risk

##        1
## 0.248699
```

So about a 25% chance. So he should get the medication, and should probably also make some changes to his lifestyle.

d

Experimenting with reducing different values in the model to normal health values produced the following results

Action	Reduction in risk
Stop smoking	3.29%
Reduce cholesterol to 195	1.49%
Reduce systolic blood pressure to 115	6.04%
Reduce BMI to 24	0.73%

The most effective actions seems to be to stop smoking and reducing the blood pressure. So the physician should tell him to a) Stop smoking and b) Excercise more and cut down on alcohol and caffeine.

e

Predicting on the test set gives the confusion matrix

```
threshold = 0.0804
pred = predict(model2b, newdata = test, type = "response")
confusion.matrix = table(test$TenYearCHD, (pred > threshold))
confusion.matrix

##
##     FALSE TRUE
##   0   287  488
##   1    15  124
```

With the confusion matrix, we can compute the metrics for our model

```
accuracy = sum(diag(confusion.matrix)) / sum(confusion.matrix)
accuracy

## [1] 0.4496718

tp = confusion.matrix[2, 2] / sum(confusion.matrix[2,])
tp

## [1] 0.8920863

fp = confusion.matrix[1, 2] / sum(confusion.matrix[1,])
fp

## [1] 0.6296774
```

For a true positive pasient, the cost will be $7500 + $165000/2.3, as the rate is decreased by a factor of 2.3. For the true negative pasients, we have no cost. For the false positive pasients we are paying for the medication, so $7500. Lastly, for the false negatives we are paying $165,000 for the disease at full rate. Our predictions on the test set therefore totals to

```
prices = matrix(c(0, 7500, 165000, 7500 + 165000 / 2.3), nrow = 2, ncol = 2, byrow=TRUE)
totalCost = sum(prices * confusion.matrix)
totalCost

## [1] 15960652
```

The pasients in the test set have a total cost of $15,960,652, an average of $17,462 per pasient.

f

The baseline model would be the current situation, which is no medication. In that case, we pay nothing for the negatives, and $165,000 for the positives.

```
nPos = sum(confusion.matrix[2,])
baselineCost = nPos * 165000
baselineCost

## [1] 22935000
```

Which gives a total economic cost of $22,935,000, and average of $25,093 per pasient.

The ideal would of course be to give all these people medication.

```
idealCost = nPos * (7500 + 165000/2.3)
idealCost

## [1] 11014239
```

Which gives a total economic cost of $11,014,239, and average of $12,051 per pasient.

If none of the pasients in the Framingham study recieves medicine, there is an economic cost of $22,935,000. However, if we somehow knew who would develop the disease and gave them medicine in advance, the total would be reduced to $11,014,239. The prediction model based on the Framingham dataset is partially able to predict who will develop the disease, and using this model would reduce the cost to $15,960,652, which is a reduction of 58.5% as compared to not prescribing medicine to anyone.

g

To calculte the area under the curve for the test set we can use

```
as.numeric(performance(rocr.pred, "auc")@y.values)

## [1] 0.7416106
```

which gives 0.7416.

Just around the part of the curve with a 0.6 false positive rate, there is a part of the curve which is horizontal, indicating that we are getting a higher false positive rate, without an increased true positive rate.

```
rocr.pred = prediction(pred, test$TenYearCHD)
plot(performance(rocr.pred, "tpr", "fpr"))
```

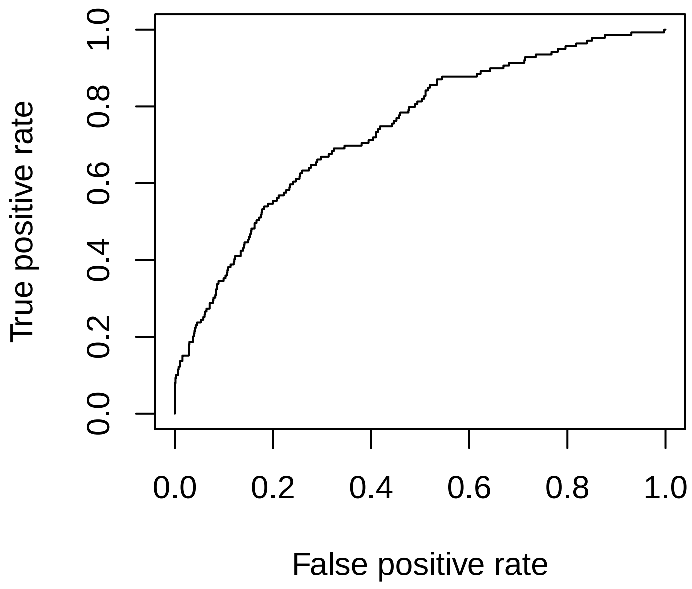

Figure 1: ROC curve for the test data of the Framingham dataset

h

In our model, both gender and age is highly significant, and their coeffisients are quite large, so let us include them. Also from the earlier discussion, reducing systolic blood pressure is a very effective action, as well as a significant variable, so let us include that as the third variable.

```
model2h = glm(TenYearCHD ~ age + male + sysBP, family = "binomial", data = train)
pred.s = predict(model2h, newdata = test, type = "response")
confusion.matrix.s = table(test$TenYearCHD, (pred.s > threshold))
totalCost.s = sum(prices * confusion.matrix.s)
totalCost.s - totalCost

## [1] -36521.74
```

This model actually results in a lower economic cost by about $36,000. Inspecting the confusion matrix reveals that this model chooses to medicate 20 more people than the original, 2 of which got the disease. So the overall accuracy is lower, but it is ”worth it” to medicate 10 people as long as at least 1 of them gets the disease.

i

I think it is very difficult and ethically questionable to estimate the monetary cost of the getting the disease. What about the relatives of the sick? Should we reduce the cost with age, as they have a lower life expectancy anyways?
If we chose to medicate some but not all, no matter how we estimate the cost of getting the disease, we will have to make some hard decisions. Like ”Your life is not worth enough to get a prescription” or ”You are not unhealthy *enough* to get a prescription”. Perhaps the analysis should only consider two choices, prescribe the medication to all or none (in which case prescribing the medicine to all would be the least costly, by a big margin).