# Breast Cancer Screening Data - Statistical Analysis

## Missing Data

My initial analysis of the data showed that all clinical teams had collected information on each characteristic except for team DN01, which had not collected information on employment status. This may have been an error in processing data after the questionnaires had been filled out, or perhaps a miscommunication in which data was to be collected, as all values from just this one site are missing. I decided to impute the results for this characteristic in team DN01 using a simple model which selected full time employment for patients above 18 years old and below 65, coinciding with leaving compulsory full time education and being able to claim a state pension respectively. Those that were outside of these were presumed to be unemployed. Although almost certainly imperfect, I thought this model would give reasonable estimates so that the data from team DN01 could still be included in the models later on.

Table 1 shows the other missing data points that were missing from the data set. The number in each cell corresponds to the amount of data points missing from the column from each site. These data points were imputed as the mean of the other points in that column so that the rows could still

| **Site** | **Missing in Height** | **Missing in Alcohol** |
| --- | --- | --- |
| 1 | 0 | 0 |
| 2 | 1 | 1 |
| 3 | 0 | 3 |
| 4 | 0 | 1 |
| 5 | 3 | 8 |
| 6 | 0 | 2 |
| 7 | 2 | 3 |
| 8 | 1 | 5 |
| 9 | 0 | 5 |
| 10 | 0 | 1 |
| 11 | 1 | 4 |
| 12 | 0 | 4 |

*Table 1: number of missing entries from each column of the dataset*

be used in the analysis. This data may be missing as the patients were unwilling to give up the information, or perhaps due to errors in processing the data post collection.

## Heterogeneity

The heterogeneity of the sample depends on factors that are unavailable from the dataset given. For example, the level of care given within each clinical team may vary depending on funding or the number of patients in total at each trial. This is unlikely to differ from a clinical team as the description notes that each data set is collected from within the same hospital. Assuming that the patients are chosen at random in a suitable form from the same population of breast cancer patients, there is no need to assess the distributions of characteristics between the clinical teams. This gives a degree of confidence in a high degree of heterogeneity between the clinical teams and so I have concatenated each team’s data into one dataset when fitting a statistical model.

## Statistical Analysis

The model I have used to fit the data is a logistic regression one which assumes that the number of scans accepted follows a binomial proportion distribution. This means that in a sample of n trials, there is an assumed probability, p, that each of the trials will be a success. In this case a “success” is a trial which results in a patient undergoing a scan. Also, the log odds[[1]](#footnote-0) of a patient accepting a screening is constant since I have assumed that each clinical team’s observations are drawn from the same population and are independent, there is no need to group and weight the data from each clinical team. It is assumed that each observation is independent.

Before the data was passed into any model, categorical variables were passed through the factor function to ensure that the R software treats them as such in the model. The initial model used all variables collected with interaction terms for alcohol, height and weight before being reduced to the final model shown in table 2. Interaction terms assess the effect on log odds of, for example, weighing a lot as well as having a high alcohol intake compared to just one or the other. I thought these interaction terms may have been statistically significant as they all may affect how healthy a patient sees themselves, which may affect how likely they are to accept a screening. All interactions could not be assessed as the steps function in R as it contained too many parameters and did not converge to a specific model. The steps function was used to reduce the initial model to the final one. This systematically adds and removes variables from the model until the model with the combination of coefficients with the lowest AIC value is found. AIC is a way of scoring how well a model minimises mean error when fitting the given data while also penalising models that have too many terms. As shown in Table 2, the model found to have the lowest AIC score in my analysis predicted the log odds of a patient taking a screening using weight, employment status and number of dependents.

| Coefficient | Estimate | P value |
| --- | --- | --- |
| Intercept | -2.17 | 0.00173 |
| Weight | 0.0372 | 3.04x10-5 |
| Part time worker\* | -1.22 | 0.00149 |
| Retired/unemployed\* | 1.77 | 9.93x10-13 |
| 1 dependent\* | -0.983 | 0.00108 |
| 2 dependents\* | -2.019 | 2.63x10-11 |
| 3 dependents\* | -3.36 | < 2x10-16 |

*Table 2 Table showing the coefficient name, estimate and p value[[2]](#footnote-1). Results given to 3 sig. fig.*

In Table 2, the estimate is the estimated change in log odds of the probability of a patient accepting a screening if a given coefficient changes by one. Starred values correspond to dummy variables that can be either be 1 or zero depending on if they are true or false. For example, the log odds of a patient in the model would differ by 0.0372 for each kilogram gained. Whereas a patient who changes from having one to two dependents would decrease their log odds by 1.04.

## Clinical Implications & Limitations of Data

To check the goodness of fit of this model, residual plots for each explanatory variable in the model were used and are shown in figure 1. These show the difference between the predicted and observed outcome at each datapoint. If the model is correct, one expects to see a trend line horizontal along the x axis with points randomly scattered above and below it. As figure 1 shows, there is some evidence that the model does not correctly utilise the age, weight and possibly also height data.

Diagram, engineering drawing

Description automatically generated

*Figure 1 residual plot of each explanatory variable in the model.*

Following this, as seen in the R code below, new models were fitted including some cubic terms. This lowered the AIC considerably and, as can be seen in figure 2, flattened the residual plots somewhat also. However, this model was very complex with 13 variables, only 9 of which were statistically significant which complicated guidance and advice potentially given to medical staff and patients regarding their choices regarding screening.

Diagram, engineering drawing

Description automatically generated

*Figure 2: residual plots for model with cubic age terms.*

The model shows that 3 of the 6 recorded variables have a statistically significant effect on the log odds and therefore the probability of a patient taking the opportunity for a scan. This change in probability may be useful for identifying patients who are less likely to undergo breast cancer screening. Identifying these patients may enable healthcare providers to ensure any cancers are found as early as possible by ensuring screenings are offered to those at risk. Of course, the NHS constitution must be kept in mind here as patient choice is central to this. Schemes such as this have been implemented to some extent by offering breast cancer screening to women in the age ranges 47 to 49 and 74 to 79 as well as between 35 and 49 more recently (Cancer Research UK, 2020). This model suggests that patients age has no significant impact on whether a patient is likely to accept a screening and incentives such as this may be better aimed at people with changing employment status, family situation or perhaps even weight.

# Bibliography

Cancer Research UK. (2020, 09 03). *Breast Screening*. Retrieved from Cancer Reaserch UK: https://www.cancerresearchuk.org/about-cancer/breast-cancer/getting-diagnosed/screening/breast-screening

1. log odds of an event is the probability that an event happens divided by the odds that it doesn’t and then transformed by the log function. Therefore the higher the log odds, the higher the probability of an event. [↑](#footnote-ref-0)
2. smaller the p value, the smaller chance that the coefficient in the true population statistically significant as opposed to the results being a fluke. A p value of 0.05 or lower is generally thought of as "statistically significant". [↑](#footnote-ref-1)