An Analysis Of Smoking Cessation

E. Webb (34683518), H. Willett (34679103), R. Lloyd-Parry (34822852),
W. Mercer (37515365) and S. Collins (34846182)

***Abstract*—The anti-smoking movement has created a cultural shift in recent decades, resulting in a rise of health concerns and relevant statistical analysis. We aim to answer the question of which path to smoking cessation based on lifestyle choices is the most successful. Throughout this report we have analysed data consisting of 4 categorical variables, and discovered that "years smoked" and "methods" are the most impactful in relation to smoking cessation. Furthermore, the best method out of the five options is individual therapy; the only one with a significant confidence interval at the 5% significance level. Amongst the numerous conclusions that are drawn from this report, the most significant was that an individual is more likely to quit smoking, the shorter the amount of time they've been smoking for, with individuals being 6.45 times more likely to restart smoking if they've been smoking for 10 years compared to someone whose only been smoking for 4 years. We conclude by outlining which group of people are most at risk of restarting.**

I. Introduction

Smoking has become a major problem for healthcare services in recent years as the amount of money spent on treating smoking related ailments has reached £3.6 billion per annum [1] and has put a huge strain on the NHS.

In 1950, British researchers demonstrated a clear relationship between smoking and cancer, based on a preliminary report that would later become known as the British Doctors study. The study itself was carried out on a population of doctors, and followed up for 50 years. It can be considered the first strong statistical proof of the correlation between smoking habits and many serious diseases, including lung cancer [2]. The best way to combat this proven link is to reduce the number of individuals who smoke. This can be done by discouraging non-smokers from starting and encouraging smokers to quit.

The anti-smoking movement has had a cultural impact, as smoking has become more taboo, and those who smoke no longer have the same status as they did a few decades ago. The social and health-conscious deterrents, such as pictorial health warnings (which have become compulsory in the UK since October 2008) [3], have encouraged the smoking population to give up their cigarettes, either on their own, or with aids recommended by the NHS.

This cultural shift has intrigued statisticians and epidemiologists, resulting in longitudinal data analyses based on smoking cessation and the variables which may explain success or failure in giving up cigarettes for good.

We look at a sample of smoking cessation data which details both categorical and indicative variables, allowing us to visualise and predict the risk of an individual's likelihood to restart smoking. These explanatory variables include: the number of years smoked (YS), the number of cigarettes smoked per day (PD), whether the individual lives with other smokers (O) and the method used to quit smoking (M). During this trial, smokers were given options and support on which methods to use to quit; including nicotine patches or gum, attending group therapy or individual talking therapy, and practising mindfulness meditation.

The binary response variable is whether or not the individual restarted smoking, (R). In our analysis, we assume the data to be independent. However, given the dataset is from an NHS trust smoking cessation clinic it is possible that there is some dependence between observations (ie. smokers in the same household may both go to this clinic affecting each other).

In this report we model the data and perform various tests in order to determine the most successful cessation method out of the 5 techniques provided in the study.

II. Methods

A. *Preliminary Analysis*

We begin our analysis by exploring the "Smoking" dataset. We first use `R` to produce a series of simple numerical summaries and graphical images in order to assess the key features of the distribution of the data. This enables us to visualise the local concentration, dispersion, shape of the data and variable relationships.

Before fitting our models, it is important to examine the relationship between the dependent and explanatory variables to gain insight into which explanatory variables are likely to be important in describing the variability in the responses. Similarly, we investigate how the explanatory variables relate to one-another to identify any potential collinearity. Since all of our explanatory variables are provided in categorical form, we use the Chi-Squared Test to assess the dependence. For a significant result, we would expect to see a Chi-Squared Statistic greater than the Critical Value for the respective degrees of freedom at a 5% significance level. Our 4 categorical explanatory variables have the following levels: $\text{YS} = 3$ (short-term/medium-term/long-term), $\text{PD} = 3$ (low/medium/high), $\text{O} = 2$ (Yes/No), $\text{M} = 5$ (see introduction).

In the event where there is an association between our categorical variables, a Chi-Squared Test provides no information as to what this relationship may be. As a result, we succeed these tests by implementing a set of odds ratios (OR) to measure the chances of one event happening relative to the chance of another event. First we calculate the odds for restarting for each level in the categorical explanatory variables. From this we derive the odds ratios for restarting to compare the likelihoods of restarting between the different levels. Furthermore, 95% confidence intervals (CI) are then constructed to see if the differences noted in the comparisons are significant. We do this in order to ascertain if the odds ratio is deviating from 1 due to an association or due to sampling variation. A significant result (i.e variation due to association) is a confidence interval that does not contain 1.

B. Model Fitting

Now that we have an intrinsic understanding of our variables, we begin fitting a model for the Smoking data.

Our data comprises of 4 explanatory variables and a response variable; the latter being whether someone restarts smoking. As the response has binary outcomes, "yes" and "no", and we have n independent and identically distributed trials, we use the Binomial family to model our variable.

The existence of 4 explanatory variables, results in there being $2^4 = 16$ potential models to choose from to represent our data. Processing and analysing each model separately is clearly impractical so we use forward selection to determine the most appropriate.

The procedure of forward selection begins with the null model, which in our case is our binary restart variable along with the intercept $(\text{logit}\,(\mu_i) = \beta_0)$. We independently add each of our 4 explanatory variables into the model and assess whether the fit improves.

If the deviance of our model is large in comparison to the χ^2_{df} critical value, then we reject the null model in favour of the more complex model. Therefore, after one iteration we will have four new models to choose from. If more than one complex model passes our test, then we select the model that best reduces the residual deviance.

Since our YS and PD variables are given in both numerical and categorical form, we first fit each variation of these and test against the null. We get that for both variables, the residual deviance is smaller for the numerical version, and so we proceed to use these for the remainder of the model fitting process.

We continue our iterations until no new predictors can be added. Once we have our final model, we add the interactions between the explanatory variables, and perform another set of deviance tests to assess the significance of the interaction model.

After our process of forward selection we are able to move on to model testing.

C. Model Testing

To ensure that our model is indeed a good one for our data, we construct a 95% confidence interval for each of the coefficients to assess the significant components of the model:

$$\left(\hat{\beta}_k - z_{1-\frac{\alpha}{2}} \times std\left(\hat{\beta}_k\right), \quad \hat{\beta}_k + z_{1-\frac{\alpha}{2}} \times std\left(\hat{\beta}_k\right)\right)$$

If this interval does not contain 0, then the given explanatory variable is a significant component in the model at the 5% significance level.

Since our model produces some insignificant coefficients, we therefore adjust the methods variable to make all the coefficients significant. We create different methods variables by grouping some of the methods together into an 'alternatives level'. We also test models with an intercept coefficient to see the effect on the model and if it helps with model interpretability. To ensure that changing the levels of the methods variable does not affect model choice, a forward model selection method was conducted. After this, we produce further confidence intervals to assess these new models.

With a final model chosen, we look at the deviance to ensure that it fits the variability of the data. We then plot the residuals on a histogram to assess their distribution and compare our residuals against a $\text{Normal}(0, 1)$ distribution by plotting the quantiles of our sample data vs the quantiles of a $\text{Normal}(0, 1)$ distribution. We also decide to plot our model on top of our data to give us an overall view, before drawing some final inference from our model.

III. Results

A. Preliminary Analysis

We plot histograms of the YS and PD and see that both variables are right skewed. For our categorical explanatory variables, we plot bar-charts with three bars per group: no. of smokers who did not restart smoking $(\bar{\text{R}})$, no. of smokers who did restart smoking (R) and the total no. of smokers in the group. We plot these three bars as it makes it easy to compare the no. of smokers in each group against the total no. in each group, and highlights significant differences in these numbers. An example of one of the four bar charts plotted is shown below.

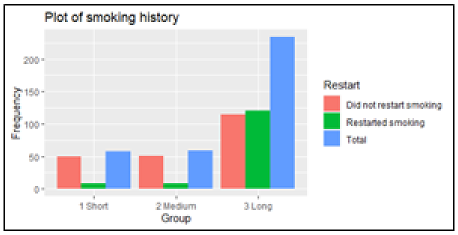

Fig. 1. A Plot Of Categorical Smoking History Against Frequency Of Restarting

In this bar-chart we see that for the short-term YS $(\text{YS} \leq 4)$ and medium-term $(4 < \text{YS} \leq 10)$ groups, the no. of smokers who did not restart smoking is much greater than the no. of smokers who did restart smoking. We further confirm our intuitions by calculating the odds for each of these groups and see that the short-term and medium-term group have odds of 0.163 (3sf) and 0.157 (3sf) respectively, while the long-term group has odds of 1.05 (3sf). Odds of 1 suggests that the likelihood of restarting is roughly equal to the likelihood of not restarting, hence we note that a significant difference in the odds highlights the effect YS has on restarting.

In our other 3 graphs (PD, M and O) we find similar relationships between groups. However, one drawback of these barcharts is that it is difficult to see which explanatory variables are significant enough to indicate if a smoker is likely to restart, and therefore if the explanatory variable is useful to include in a model. Following this, we conduct Pearson's Chi-Squared Tests for independence between the categorical explanatory variables YS, PD, M, O and against the response variable R.

The Chi-Squared Statistics and Critical Values can be seen in Table I. When the Chi-Squared Statistic was tested against the Critical Values it was concluded that there was dependence between the explanatory variable and the response variable in all cases but the O variable. When testing between the explanatory variables, just YS and PD shared a dependent relationship.

Variable 1	**Variable 2**	**Chi-Squared Statistic (3 s.f.)**	**Critical Value (to 3 s.f.)**	**Significant?**
YS	R	45.9	$5.99(\text{df} = 2)$	Yes
PD	R	10.7	$5.99(\text{df} = 2)$	Yes
M	R	33.2	$9.49(\text{df} = 4)$	Yes
O	R	0.542	$3.84(\text{df} = 1)$	No
YS	PD	11.8	$9.49(\text{df} = 4)$	Yes
YS	M	2.11	$15.5(\text{df} = 8)$	No
YS	O	0.831	$5.99(\text{df} = 2)$	No
PD	M	6.90	$15.5(\text{df} = 8)$	No
PD	O	0.884	$5.99(\text{df} = 2)$	No
M	O	4.67	$9.49(\text{df} = 4)$	No

TABLE I
CHI-SQUARED TEST RESULTS

Having identified the dependence between variables, we now examine the nature of this relationship. Table II below shows the significant relationships and differences for restarting within the categorical variables' levels - variable level 1 relative to variable level 2.

Variable 1	**Variable 2**	**OR (3.s.f)**	**95% CI (3.s.f)**
Smoked $>$ 10 Yrs	Smoked $<$ 10 Yrs	6.45	$(2.18, 19.1)$
Smoked $>$ 10 Yrs	Smoked $4 < x < 10$ Yrs	6.71	$(2.26, 19.9)$
High Restart Freq.	Low Restart Freq.	2.68	$(1.35, 5.30)$
Group Support	Individual Support	5.39	$(1.58, 18.4)$
Meditation	Individual Support	6.15	$(1.80, 21.0)$

TABLE II
SIGNIFICANT ODDS RATIOS

From this we draw a number of conclusions.

- People are 6.45 times more likely to restart if they have smoked for greater than 10 years compared to less than 4 years.
- People are 6.71 times more likely to restart if they have smoked for greater than 10 years compared to 4-10 years.
- People are 2.68 times more likely to restart if they smoked a high amount compared to a low amount.
- People are 5.39 times more likely to restart if they have group support compared to individual support.
- People are 6.15 times more likely to restart if they have meditation compared to individual support.

B. *Model Fitting*

We first note that for both YS and PD variables the numerical versions have a lower residual deviance, and therefore we will proceed with those versions in our future iterations.

In our first iteration (see Table III), we have that all of our models pass except $\text{logit}(\mu_i) = \beta_0 + \beta_1 \text{O}_i$, and therefore it is the only model in which the null fits our data better. Our model with the lowest residual deviance is $\text{logit}(\mu_i) = \beta_0 + \beta_1 \text{YS}_i$ so this becomes our new null model for the second iteration.

Model ($\text{logit}(\mu_i) =$)	**Deviance**	**Test Statistic**	**Critical Value**	**Accept Reject H0**
Null: $= \beta_0$	467.67	-	-	-
$= \beta_0 + \beta_1 \text{M}_i$	431.13	36.55	9.49	Reject
$= \beta_0 + \beta_1 \text{O}_i$	467.13	0.54	3.84	Accept
$= \beta_0 + \beta_1 \text{YS}_i$	376.18	91.49	3.84	Reject
$= \beta_0 + \beta_1 \text{PD}_i$	454.58	13.09	3.84	Reject

TABLE III
1ST ITERATION

In our second iteration (see Table IV), we note that only one of our new models that has a significant test result: $\text{logit}(\mu_i) = 1 + \text{YS}_i + \text{M}_i$. Therefore, this is our immediate choice for our new null model.

Model ($\text{logit}(\mu_i) =$)	**Deviance**	**Test Statistic**	**Critical Value**	**Accept/ Reject H0**
Null: $= \beta_0 + \beta_1 \text{YS}_i$	376.18	-	-	-
$= \beta_0 + \beta_1 \text{YS}_i + \beta_2 \text{M}_i$	333.61	42.57	9.49	Reject
$= \beta_0 + \beta_1 \text{YS}_i + \beta_2 \text{PD}_i$	373.98	2.20	3.84	Accept
$= \beta_0 + \beta_1 \text{YS}_i + \beta_2 \text{O}_i$	375.97	0.21	3.84	Accept

TABLE IV
2ND ITERATION

In our third iteration (see Table V), we have the result that none of our new models fit the data better than our $\text{logit}(\mu_i) = \beta_0 + \beta_1 \text{YS}_i + \beta_2 \text{M}_i$ model and so we conclude that this is our best model for our data prior to testing for interaction.

Model ($\text{logit}(\mu_i) =$)	**Deviance**	**Test Statistic**	**Critical Value**	**Accept/ Reject H0**
Null: $= \beta_0 + \beta_1 \text{YS}_i + \beta_2 \text{M}_i$	333.61	-	-	-
$= \beta_0 + \beta_1 \text{YS}_i + \beta_2 \text{M}_i + \beta_3 \text{PD}_i$	330.56	1.68	3.84	Accept
$= \beta_0 + \beta_1 \text{YS}_i + \beta_2 \text{M}_i + \beta_3 \text{O}_i$	332.85	0.76	3.84	Accept

TABLE V
3RD ITERATION

Table VI shows the result for our interaction model tests. We conclude that we are no better off using interaction in our model, therefore our latest null model $\text{logit}(\mu_i) = \beta_0 + \beta_1 \text{YS}_i + \beta_2 \text{M}_i$ is the best fit for our data.

Model	**Deviance**	**Test Statistic**	**Critical Value**	**Accept/ Reject H0**
Null: $\beta_0 + \beta_1 \text{YS}_i + \beta_2 \text{M}_i$	333.61	-	-	-
$\beta_0 + \beta_1 \text{YS}_i + \beta_2 \text{M}_i + \beta_3 \text{YS}_i \text{M}_i$	327.98	5.63	9.49	Accept

TABLE VI
TESTING FOR INTERACTION

C. Model Testing

The results of the coefficient confidence intervals of $\operatorname{logit}(\mu_i) = \beta_0 + \beta_1 \mathrm{YS}_i + \beta_2 \mathrm{M}_i$ are displayed in Table VII.

Coefficient	95% CI	Conclusion
Intercept (β_0)	$(-3.08, \quad -1.55)$	Significant
β_1	$(0.09, \quad 0.16)$	Significant
β_2:meditation	$(-0.77, \quad 1.06)$	Insignificant
β_2:indiv	$(-2.78, \quad -1.07)$	Significant
β_2:gum	$(-0.42, \quad 1.48)$	Insignificant
β_2:patches	$(-0.5, \quad 0.92)$	Insignificant

TABLE VII
CONFIDENCE INTERVAL FOR COEFFICIENTS AND LEVELS OF ORIGINAL MODEL

This shows that the coefficients for the methods of quitting smoking: meditation, gum and patches are all not significant.

We now alter the M variable levels to produce a set of new potential categorical variables for the methods used. We call this 'Alternatives' (A). To save space, the test results for this have been omitted.

When choosing our final model we look back at our preliminary analysis. The odds values for each of the methods suggest that the most significant method for preventing restarting smoking is individual therapy as the log-odds is significantly negative and not close to zero. A model with an intercept was chosen as it allows the other methods of quitting smoking to act as a baseline. We therefore choose the model:

$$\operatorname{logit}(\mu_i) = \beta_0 + \beta_1 \mathrm{YS}_i + \beta_2 \mathrm{A}_i \tag{1}$$

$$\text{Where} \quad \mathrm{A}_i = \begin{cases} 1, & \text{if individual therapy} \\ 0, & \text{otherwise} \end{cases}$$

The significance test for the final model produced Table VIII:

Coefficient	Value	95% CI	Conclusion
Intercept (β_0)	-2.10	$(-2.66, \ -1.54)$	Significant
β_1	0.12	$(0.09, \ 0.15)$	Significant
β_2	-2.10	$(-2.83, \ -1.37)$	Significant

TABLE VIII
CONFIDENCE INTERVAL FOR COEFFICIENTS AND LEVELS OF NEW MODEL

The interpretation for each of the coefficients is as follows:

- β_0 is the baseline case and shows that if they are given any method for stopping other than individual, the likelihood of them restarting decreases by 2.10.
- β_1 is interpreted as: with every additional year smoked, the likelihood of an individual restarting smoking increases by 0.12.
- β_2 is interpreted as: if someone is given individual therapy, the likelihood of them restarting smoking decreases by 2.10 from any other method.

D. Model Checking

The deviance of the model is 334.84 with 347 degrees of freedom, which is less than 391.44 (the χ^2_{347} critical value at the 5% level) and so we conclude that the model is able to adequately describe the variability in the data.

We then plot the residuals on a histogram as seen below (Figure 2). We can see that this histogram is roughly normally distributed with a mean about zero which suggests that overall our model is good at predicting the likelihood of a smoker restarting given the number of years they have smoked and whether they have undergone individual therapy.

From the Q-Q plot (Figure 2) we see that there is a "banding structure" around $(0.3, -0.1)$ which we expect to see due to the use of the logistic regression function. Despite this banding structure we can see that the data roughly lies on the line $\mathrm{y} = \mathrm{x}$ giving us confidence that our model is a good descriptor of the data.

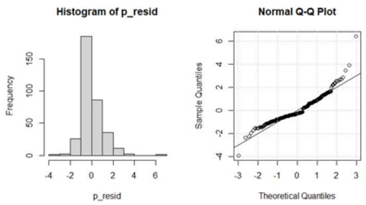

Fig. 2. Plot Of Residuals On A Histogram & Q-Q Plot Of Residuals

Our final model has the logistic function:

$$\mu_i = \frac{\exp\left(\beta_0 + \beta_1 \mathrm{YS}_i + \beta_2 \mathrm{A}_i\right)}{1 + \exp\left(\beta_0 + \beta_1 \mathrm{YS}_i + \beta_2 \mathrm{A}_i\right)} \tag{2}$$

To plot our model on top of our data, we first separate the data into smokers who have undergone individual therapy (blue) and smokers who have undergone another method of quitting smoking (green). (See Figure 3).

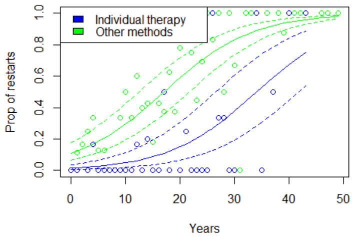

Fig. 3. Plot Of Our Model Over The Dataset

The solid lines on the graph represent our model with their coefficients at their MLEs. As we can see from these solid lines and the grouping of the data, smokers who underwent individual therapy are more likely to not restart smoking. However as we can see for both data sets the proportion of smokers who restart increases with the no. of years the person has smoked. We can also see from our graph that the proportion of restarts for smokers who underwent other methods has a larger variation. This potentially suggests that other factors like living with smokers has some effect on the proportion of smokers who restart however its role is limited in the ability to predict if a smoker is likely to restart smoking.

IV. Discussion

A. Inference

Understanding the extent to which smokers are most likely to permanently quit under different social and personal circumstances proves important on both a widespread and individual basis.

Using Equation (2), alongside the values of our coefficients, we are able to calculate some probabilities for different scenarios. From this we infer which groups are most at risk of restarting smoking.

The fitted probability that a smoker of 50 years, who receives an alternative treatment, restarts smoking is $\mu_{50,A=0} = 0.980$ (3sf). Conversely, for a smoker of the same time period, who instead receives individual therapy, the probability is just $\mu_{50,A=1} = 0.858$ (3sf). This highlights that for a 50 year old smoker, individual therapy reduces the probability of restarting by 12.2%. An individual that has smoked for just 1 year, and has individual therapy, has a $\mu_{1,A=1} = 0.017$ (3sf) probability of restarting, emphasising the effect the no. years smoked has on restarting - in this case 84.1%.

The no. of years at which the probability of restarting, for someone with alternative treatment, overtakes $\mu = 0.5$ is 18. This can be improved to 35 years if the person has individual therapy. Therefore from our analysis, we would define an at risk individual as someone that has smoked for over 35 years and would strongly recommend that all patients receive individual treatment.

B. Conclusion

Despite the numerous resources offered to aid giving up smoking, we realise from our results that only individual therapy, out of the selection given, has any significance to whether someone restarts smoking again or not. This is an important finding as it uncovers the idea that patients are better off receiving individual therapy. At first glance, this sounds counter intuitive as many programs designed to cease bad habits, such as Alcoholics Anonymous, are built upon the idea of unity and group therapy. Our results help us to dispel this assumption and ascertain the best treatment for the patients.

No statistical model is without limitations. By grouping all but the individual method, into one level of the categorical variable, we have produced a more parsimonious model. However, this limits our ability to make inference, specifically about these methods on their own. Our data only considers 4 explanatory variables in smoking cessation and it is possible that there are more important variables it doesn't consider. We have no information regarding the length of restart follow-up period, and hence our inference on the best methods to prevent restarting could be time sensitive (i.e. non-restart patients still return to smoking.) Another aspect to the limitations of our work is that the data hasn't been collected from a true clinical trial and so cannot be confidently compared with previous studies and may not correlate with preceding expectations. This could cause it to be difficult to associate with real-world applications.

A potential pathway into future analysis that may improve our work could be to gather a true data set. To improve our study, we may look into other potential variables that contribute to success in smoking cessation such as age, whether the participants parents were smokers, and gender. Gender differences would be interesting to investigate as, "women may be at relatively greater risk of smoking-related diseases than men and tend to have less success than men in quitting smoking" [4]. Links have also been made between smoking and depression: "smokers with major depression [are] less successful at their attempts to quit" than other groups in the trial who don't suffer from major depression [5]. Since "the lifetime prevalence of a major depressive disorder in women (21.3%) is almost twice that in men (12.7%)", [6], it would be fascinating to explore any interactions between a depressive disorder and a variable we have already analysed such as number of years smoked, across differing genders. We could further discuss the link between depression and how successful these individuals are at quitting smoking considering the apparent statistical odds stacked against them.

We address that no model is perfect, and the ability to determine precisely the most successful pathways to smoking cessation can only ever be achieved to a certain extent. However, it remains an area of considerable scientific intrigue from both a social and health beneficial point of view.

References

[1] "Smoking and the public purse," accessed: 24/03/21. [Online]. Available: https://iea.org.uk/wp-content/uploads/2017/08/Smoking-and-the-Public-Purse.pdf

[2] R. Doll, R. Peto, J. Boreham, and I. Sutherland, "Mortality from cancer in relation to smoking: 50 years observations on british doctors," *British journal of cancer*, vol. 92, no. 3, pp. 426–429, 2005.

[3] "Smoking warning labels," accessed: 25/03/21. [Online]. Available: https://ash.org.uk/category/information-and-resources/packaging-labelling-information-and-resources/warning-labels/

[4] "Smoking cessation in women," accessed: 07/04/21. [Online]. Available: https://link.springer.com/article/10.2165/00023210-200115050-00005

[5] M. Alexander H. Glassman, M. John E. Helzer, P. Lirio S. Covey, and et al, "Smoking, smoking cessation and major depression," 1990. [Online]. Available: https://jamanetwork.com/journals/jama/article-abstract/383334

[6] "Depression in women," accessed: 07/04/21. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0026049505000363?casa$_token = iNxL2glSXkEAAAAA :$ $eESsw2bFzd3lyWTV9VNEpq251ya3SPa0TvjkRW$ $vNTzlQVacThzs0cdpTRV6GXi1yjrlpRAqt$