Understanding Factors Influencing Cancer Patient Survival: A Mixed Models Approach

MATH3092: Mixed Models Report

Shrichand Bhuria
201800797

Department of Mathematics
University of Leeds
United Kingdom
April 30, 2024

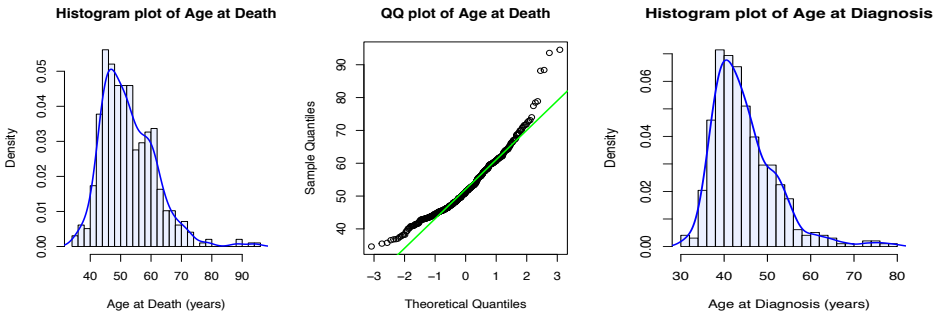
1 Introduction

Cancer presents a significant global health challenge, often complicating both diagnosis and treatment. In this study, we delve into an extensive dataset named CWsample, comprising 490 patient records sourced from twelve hospital sites. This dataset offers crucial insights into various aspects of cancer patients' profiles, including their age at death, age at diagnosis, cancer type, and demographic details. Upon initial examination, we discover that, on average, patients succumb to cancer at the age of 52.6 years. Age at diagnosis emerges as a pivotal predictor of patients' survival time. Furthermore, personal characteristics such as gender and quality of life, represented by the 'HighQualityOfLife' variable, significantly influence age at death. Our objective in this report is to validate these findings through meticulous analysis of the CWsample dataset.

The dataset includes a SiteID number, delineating the hospital each patient attended. It also encompasses variables such as 'Sex', indicating the patient's gender, and 'HighQualityOfLife', providing insights into their quality of life. Additionally, details regarding cancer type, treatment regimens, and tumor characteristics offer a comprehensive understanding of each patient's profile. Our analytical approach leverages the 'Stats' package for constructing linear regression models using the 'lm' function, along with the 'lme4' package for developing random intercept and random slope models via the 'lmer' function. Through this comprehensive analysis, we aim to unravel the intricate factors influencing age at death among cancer patients, thereby offering valuable insights for clinical practice and guiding future research endeavors.

2 Exploring Model Analysis Techniques

We find that the patient age at death varies quite greatly between patients, as shown from the histogram plot on the left in Fig. 1a. As the density plot depicts a roughly bell-shaped curve of the age of death variable, we can infer that this variable is normally distributed. In addition, the normal quantile-quantile (Q-Q) plot on the right shows little curvature, further suggesting that 'AgeDeath' follows a normal distribution. Similarly, the time at the age of diagnosis for the patient also varies between patients and appears to follow a normal distribution as illustrated in Fig. 1b.



(a) Histogram (left) and Q-Q (right) plot for the age at death of patients **(b)** Histogram plot for age at diagnosis

Fig. 1. (a) This shows the age at death variable distributed normally between patients based on model (1), and (b) this shows the distribution of the age at diagnosis variable.

When assuming the ages of death are independent and identically distributed, we propose the model:

$$y_i = \beta_0 + e_i \quad \text{with} \quad e_i \sim N(0, \sigma^2). \tag{1}$$

where y_i denotes the age of death of the i-th patient in our CWsample, such that $i = 1, 2..., 490$ and e_i denotes the random error for each individual. Using the method of ordinary least squares via the 'lm' function in R, we find an estimate for the population mean, $\hat{\beta_0} \approx 52.6$ years, which indicates the average patient age of death. Additionally, the model has an estimated variance of $\hat{\sigma_0}^2 \approx 8.8^2$, suggesting a large spread of around 9 years, signifying significant variability in the distribution of age of death between individuals. In an attempt to decrease this variance, we propose the following linear regression model:

2.1 Linear Regression Model:

Model A, which investigates the relationship between age of death and age at diagnosis:

$$\text{Model A:} \quad y_i = \beta_0 + \beta_1 x_{1i} + e_i \quad \text{with} \quad e_i \sim N(0, \sigma^2). \tag{2}$$

Here x_{1i} denotes the time at the age of diagnosis, in years. Defining this model in R using the 'lm' function, the estimates of the parameters and variance become $\hat{\beta_0} \approx 7.45$ years, $\hat{\beta_1} \approx 1.01$ years, and $\hat{\sigma_0}^2 \approx 4.63^2$, respectively. As the variance parameter has decreased from the previous model and the coefficient of β_1 is approximately one, there is evidence to suggest that the individual's age of death increases at the same rate as the age of diagnosis.

The scatter plot on the left of Fig. 2 shows that the points are equally spread above and below the purple regression line, suggesting that the linear regression model is a good fit to the data. The positive gradient of this line suggests a positive correlation between a patient's age at death and age at diagnosis,

as demonstrated by the positive coefficient β_1. However, the Q-Q plot on the right of Fig. 2 shows a slight curvature, with many points lying above the straight orange line. This suggests that the distribution of age at death may not support the normality assumption in this linear regression model. To improve this model, we shall consider whether the different hospital sites where patients were treated have an influence on the relationship between age at death and age at diagnosis. The number of patients from each of the 12 hospital sites is summarized in Table 1 below.

SiteID	**4**	**5**	**7**	**8**	**10**	**15**	**16**	**18**	**19**	**20**	**25**	**31**
Number of Patients	153	12	20	114	52	6	19	27	6	37	14	30

Table 1: Summary of the number of patients from each hospital site.

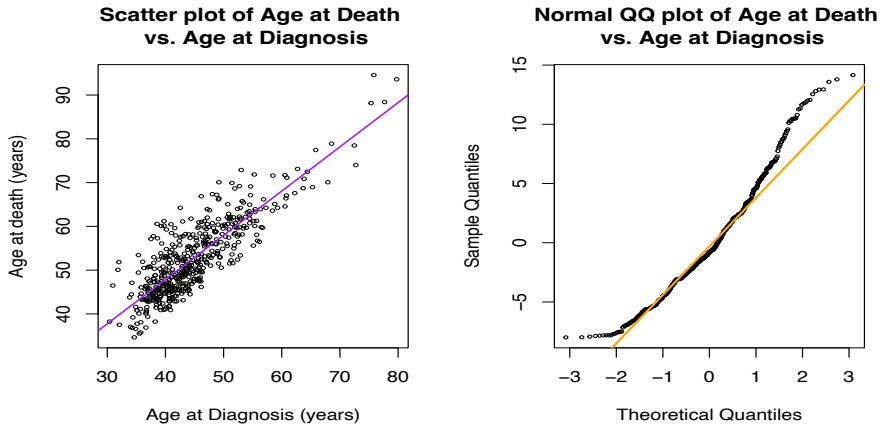


Fig. 2. Plot of age at death against the age at diagnosis, with a purple estimated regression line based on our linear regression Model A (left) and the corresponding normal Q-Q plot of the estimated error residuals for this model (right), with an orange Q-Q line added to the plot.

2.2 Exploring Additional Covariates

In Model A, which considers only Age at Diagnosis as the dependent variable, there may be limitations in capturing the full complexity of patient outcomes, particularly when patients diagnosed at the same age exhibit varying characteristics. To enhance the predictive ability of our model, we investigate the impact of additional patient characteristics. Specifically, we consider three potential covariates: Sex, High Quality of Life (HighQualityOfLife), and Stage3or4 Status. To inform our selection, we examine box plots illustrating the distribution of Age at Death across different categories for each covariate, as shown in Figure 3.

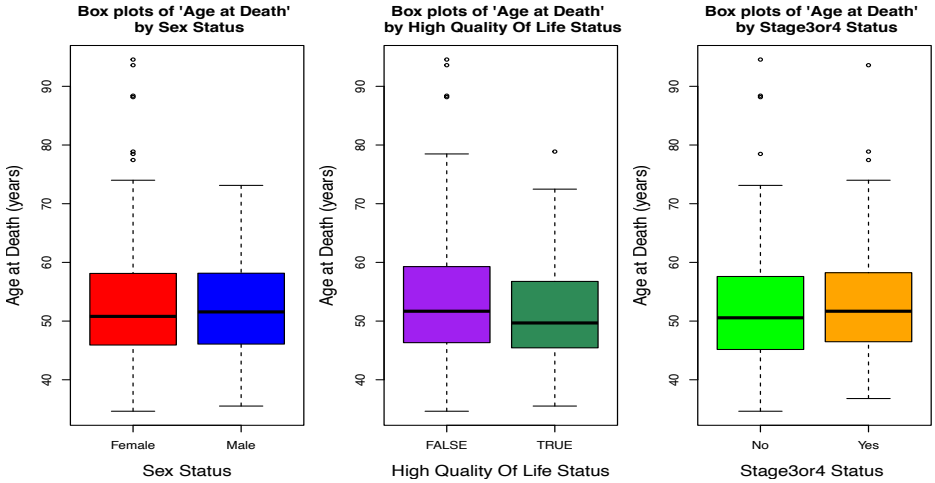


Fig. 3. Box plots illustrating the distribution of Age at Death across various categories of Sex, High Quality of Life, and Stage3or4 Status.

In these plots, we observe variations in the mean age of death across different categories of each covariate. Specifically, there are noticeable differences between those with high quality of life (seagreen box) and those

without (purple box), in comparison to both other covariates. To objectively determine which covariate has the strongest influence on age at death, we conduct separate linear regression analyses for each covariate. By examining the p-values obtained from these analyses, we can assess the significance of each covariate. Upon comparison, High Quality of Life emerges as the covariate with the smallest p-value ($p = 0.01531$), followed by Sex Status ($p = 7158$) and Stage3or4 ($p = 0.128$). Therefore, we prioritize High Quality of Life as the covariate to include in our primary linear regression Model A due to its stronger association with patients' age at death.

In the context of our analysis, we extended Model A by introducing a new covariate, 'HighQualityOfLife', to investigate its effect on the age of death alongside the age at diagnosis of patients. Model B is defined as:

$$\text{Model B:} \quad y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i \quad \text{with} \quad e_i \sim N(0, \sigma^2), \tag{3}$$

where x_{2i} represents the High Quality of Life status of patient i, with '1' indicating having High Quality of Life and '0' indicating non-High Quality of Life. Using the 'lm' function, the estimated parameters and variance for Model B are approximately $\hat{\beta}_0 \approx 7.83$ years, $\hat{\beta}_1 \approx 1.0067$ years, $\hat{\beta}_2 \approx -0.59$, and $\hat{\sigma_0}^2 \approx 4.63^2$, respectively. The coefficient of approximately -0.59 for High Quality of Life in Model B indicates that having High Quality of Life is associated with a decrease in the expected age of death by about 0.59 years compared to not having High Quality of Life, holding all other factors constant. This finding suggests that non-High Quality of Life status has a significant impact on survival age, with non-High Quality of Life patients generally expected to live slightly longer, on average, compared to High Quality of Life patients, even when diagnosed at the same age.

2.3 Random Intercept Model

In Fig. 4, we observe that the variance around the estimated mean of age at death from the original model greatly varies between patients in different hospital sites. For instance, all patients in SiteID 7 fall below the estimated mean, while in SiteID 15, all patients fall above the estimated mean. This suggests that a multilevel model that takes into account the patient's hospital site would provide a better fit to the data than the previous models. The influence of clustering on the patient's age at death can be more clearly seen in Fig. 5.

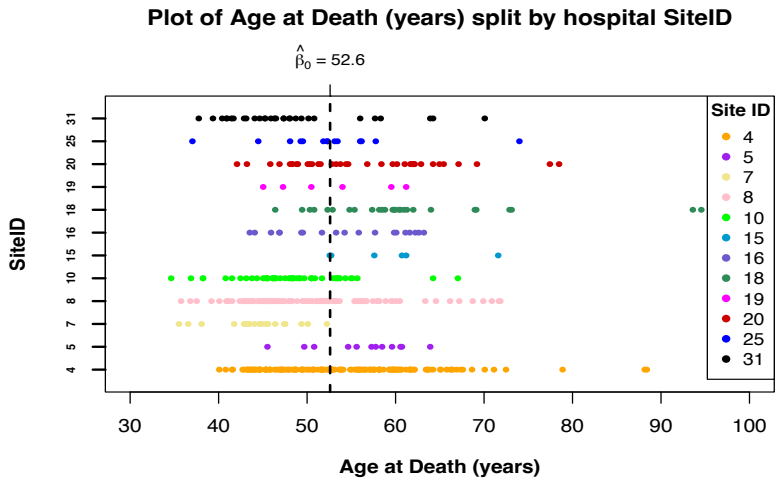


Fig. 4. Plot of patient age at death, grouped by the hospital sites they were treated in, with a dotted line at the estimated mean of the original model.

In this analysis, we expanded our previous model (2) to see if the hospital where patients were treated affects their age at death. We introduced a new model called the random intercepts model (Model C), which helps us understand how the relationship between a patient's age at diagnosis and age at death varies across different hospital sites. In Model C, we define the age of death of the i-th patient in the j-th hospital site as y_{ij}. This model assumes that the age of death of patients is independent and identically distributed. The equation for Model C looks like this:

$$\text{Model C:} \quad y_{ij} = \beta_0 + \beta_1 x_{1ij} + u_{0j} + e_{0ij} \quad \text{with} \quad u_{0j} \sim N(0, \sigma^2_{u0}) \quad \& \quad e_{0ij} \sim N(0, \sigma^2_{e0}). \tag{4}$$

Here, x_{1ij} represents the age of diagnosis of the i-th patient at the j-th hospital site and u_{0j} represents the random effect on hospital sites, capturing the variability in age of death between different hospital sites. After fitting Model C in R using the 'lmer' function, we found estimated parameters $\hat{\beta}_0$ to be around 10.71 and $\hat{\beta}_1$ to be approximately 0.948. The variance estimates are $\hat{\sigma}^2_{u0}$ at about 2.58 squared years at the hospital site level and $\hat{\sigma}^2_{e0}$ at approximately 4.17 squared years at the patient level.

We also calculated the Intraclass Correlation Coefficient (ICC) to quantify the proportion of variance attributable to differences between hospital sites. The ICC value obtained was approximately 0.277, indicating that about 27.7% of the total variance in age of death can be attributed to differences between hospital sites. In the scatter plot shown in Fig. 5, it's evident that the age of death varies between hospital sites, as indicated by the considerable variation in the intercepts of the dashed lines. This variation suggests clustering by hospital SiteID.

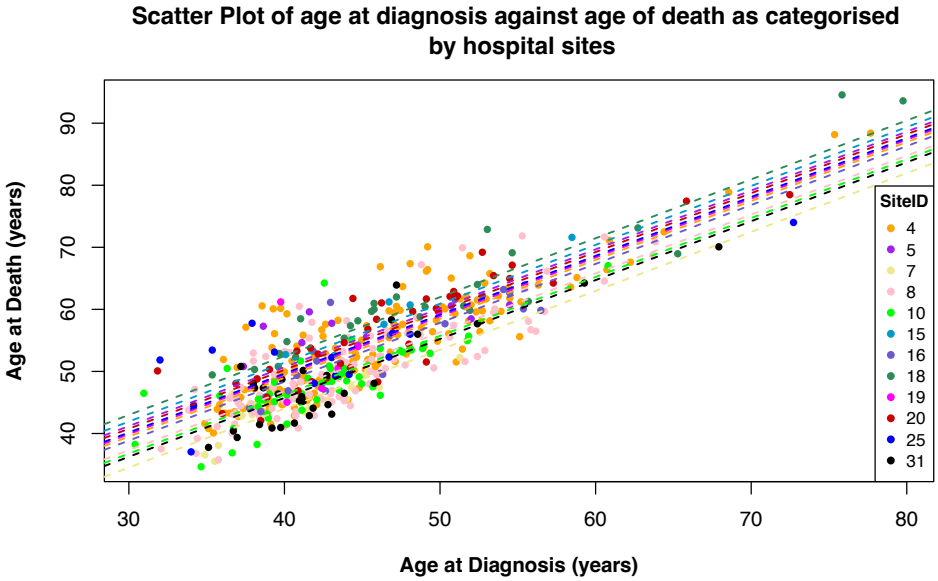


Fig. 5. Scatter plot of patient's age at death against age at diagnosis, with dashed colored lines representing regression lines corresponding to hospital sites from the random intercept model.

Additionally, we can evaluate the assumption of normality for the residuals by examining the normal Q-Q plots, as depicted in Figure 6. In the right plot, representing the level 2 residuals, the points closely align with the purple line, suggesting that these residuals approximately conform to a normal distribution. However, in the left plot, illustrating the level 1 residuals, we observe a slight deviation from the red line. This indicates some uncertainty regarding whether these residuals adhere strictly to a normal distribution.

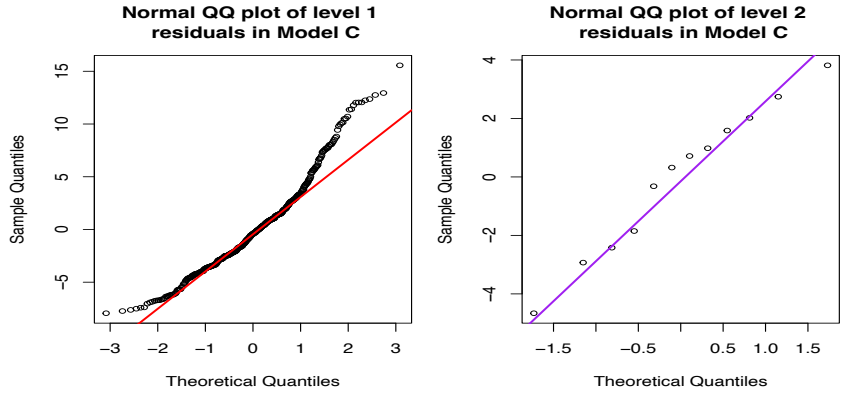


Fig. 6. Normal Q-Q plots for level 1 residuals (left), fitted with a red Q-Q line, and level 2 residuals (right), fitted with a purple Q-Q line for Model C.

Likelihood Ratio Test: We perform a likelihood ratio test to determine whether there is enough evidence to prefer the multi-level Model C over the single-level linear regression Model A. Since the only additional parameter in Model B is the level two variance parameter, σ^2_{u0}, we test the following hypotheses:

$$H_0 : \sigma^2_{u0} = 0 \quad \text{vs.} \quad H_1 : \sigma^2_{u0} > 0.$$

Conducting this test yields a p-value of $p = 2.2 \times 10^{-16}$ at a significance level of 5%. It suggests that the observed data is unlikely to have occurred under the assumption of the null hypothesis (in this case, that the variance parameter σ^2_{u0} is zero). Therefore, we reject the null hypothesis in favor of the alternative hypothesis, indicating that there is sufficient evidence to support the superiority of Model C over Model A.

Interpretation: The significant p-value suggests that including the random intercept for SiteID in Model C improves the model fit compared to Model A. Therefore, considering clustering by hospital site has a significant impact on explaining the variability in the age of death beyond what is explained by the age of diagnosis alone. Hence, Model C is preferred over Model A due to its better fit and ability to capture the clustering effect of hospital sites.

2.4 Enhanced Model Incorporating Random Slopes:

In our analysis, we consider a scenario where patients' age at death not only depends on individual characteristics but also varies across different hospital sites. To capture this variability, we extend Model B by introducing a random slope model, denoted as Model D:

$$\text{Model D:} \quad y_{ij} = \beta_0 + \beta_1 x_{1ij} + u_{1j} x_{1ij} + \beta_2 x_{2ij} + u_{0j} + e_{0ij} \tag{5}$$

Here, x_{1ij} denotes the age at diagnosis of patient i at hospital site j, while x_{2ij} indicates the covariate representing the patient's high quality of life (TRUE for patients with high quality of life). We found estimated parameters: $\hat{\beta}_0$ around 10.67, $\hat{\beta}_1$ approximately 0.944, and $\hat{\beta}_2$ approximately 0.092. The variance estimates are: $\hat{\sigma}^2_{u0}$ at about 0.3255 squared years at the hospital site level, $\hat{\sigma}^2_{e0}$ at approximately 4.18 squared years at the patient level, and random slope variance $\hat{\sigma}^2_{u1}$ at 0.0651.

Model D demonstrates significant improvement over Model B at the 5% significance level, with a p-value of 8.326673×10^{-16}, calculated using adjusted p-values considering 1 additional covariate and 1 additional variance parameter. The regression lines for Model D over the scatter plot of age at death versus age at diagnosis are depicted in Fig. 7. Since Model D incorporates random slopes, the regression lines exhibit variations in both intercepts and slopes, contingent on the hospital site where patients were treated. Additionally, slight variations are observed between patients with high quality of life and those without. Notably, patients from hospital sites '7' and '31' exhibit lower ages at death, while those from '18' and '15' tend to have higher ages at death. This pattern resonates with the findings from Model B, further reinforcing our conclusions.

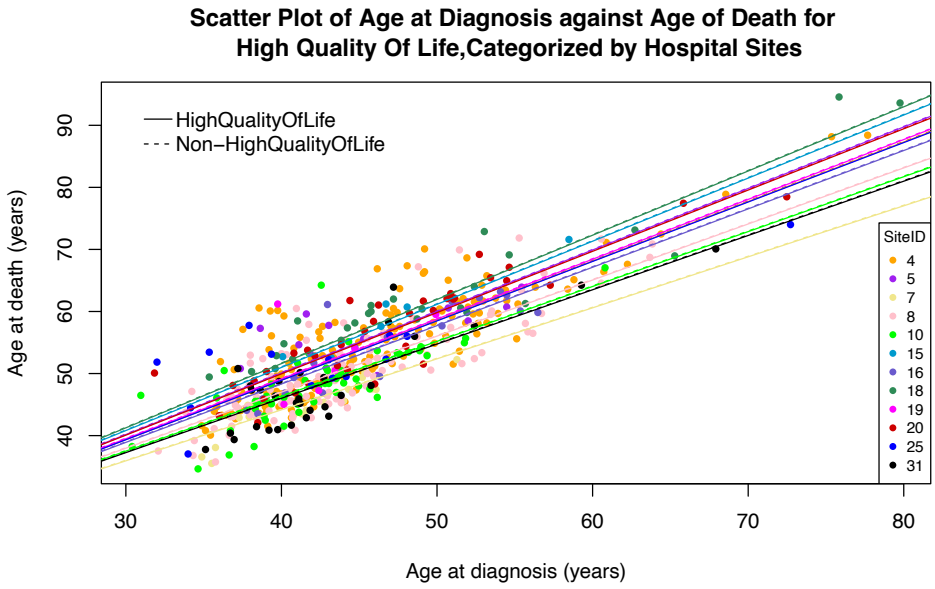


Fig. 7. Plot of patients' age at death against age at diagnosis. The colored lines represent regression lines corresponding to the hospital sites where the patients were treated. Dashed lines indicate non-high quality of life patients, while solid lines represent high quality of life patients.

In Fig. 8, we explore how the clustering of patients within hospital sites affects their age at death using caterpillar plots presenting the level one and level two residuals derived from Model D. The left plot, showcasing the 'Intercept Residuals', reveals the variations in intercepts across different hospital sites compared to the overall population intercept. Notably, distinct clusters emerge, particularly evident with hospital sites '7' and '18', which significantly deviate from the expected pattern represented by the green line. These residuals' intervals exclude zero, indicating notable differences between these hospital sites and the average regarding patients' age at death and racial demographics. Similarly, the right plot illustrates the 'AgeDiagnosis' Residuals, indicating deviations in the slopes (coefficients for the 'AgeDiagnosis' variable) for each hospital site from the overall population slope. Here, we again observe pronounced deviations for hospital sites '7' and '18', although in opposing directions compared to the left plot. Furthermore, both plots highlight wide intervals for most hospital sites, indicating considerable uncertainty in estimating these sites' effects on patients' age at death. This uncertainty is especially notable for hospital sites '15' and '19'. Therefore, cautious interpretation of the clustering effects on patients' age at death is essential, emphasizing the need for further investigation to understand the underlying factors contributing to this variability.

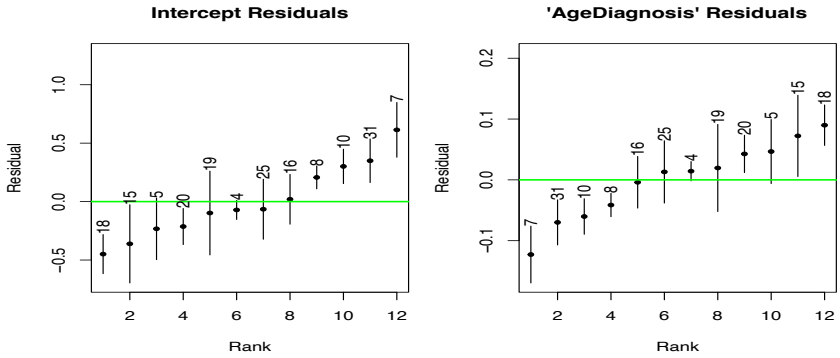


Fig. 8. Caterpillar plots illustrating the level one and level two residuals for Model D, with a 95% confidence interval. The horizontal green lines indicate the average state, corresponding to a residual of zero.

Despite the apparently good fit of the model shown in Fig. 7 and 8, it appears that using age at diagnosis as a random effect introduces complexity, and the variance component associated with it is nearly zero. This is evident from the normal Q-Q plot of these residuals in Fig. 9, where the sample quantities on the y-axis exhibit a very narrow range between $(-0.10, 0.05)$.

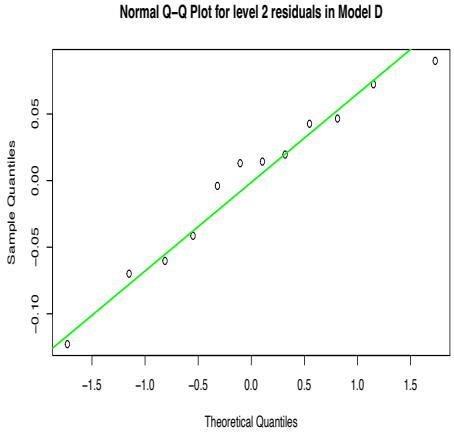


Fig. 9. Normal Q-Q plot of level 2 residuals for Model D, with a green Q-Q line.

3 Discussion

In summary, our analysis of the CWsample dataset reveals that the average age at death for patients is approximately 52.6 years. Moreover, we observe a positive correlation between a patient's age at diagnosis and their age at death. Through likelihood ratio tests, we ascertain with a 95% confidence level that the hospital site where a patient is treated significantly influences the relationship between age at diagnosis and age at death. Specifically, patients treated at hospital sites such as '18' and '15' tend to have higher ages at death, while those from sites '7' and '31' exhibit lower ages at death. This variation could potentially be attributed to differences in resources and funding among hospital sites, leading to disparities in treatment effectiveness and patient outcomes. Additionally, our analysis highlights the substantial impact of high quality of life on patients' age at death. The scatter plots, fitted with regression lines for each hospital site, demonstrate the influential role of site clustering on patients' age at death, as evidenced by models B and D. These findings underscore the importance of considering hospital site as a significant factor in predicting patient outcomes and emphasize the need for targeted interventions and resource allocation to improve healthcare delivery across different sites.

Despite encountering challenges in model convergence due to limited data availability, we managed to develop models that provide valuable insights into factors influencing cancer patient survival. While our models generally exhibit good fit, the assumption of normality for residuals, particularly at level 2, may be violated. This is evident from the narrow range observed in the normal Q-Q plot of level 2 residuals in Model D, indicating potential complexities in the choice of random effects. These findings highlight the importance of cautious interpretation and further validation of results in real-life clinical settings. These conclusions have potential real-life relevance, particularly in informing clinical practices and healthcare policies aimed at improving cancer patient outcomes. Understanding the interplay between age at diagnosis, hospital characteristics, and patient demographics can aid in the development of targeted interventions to enhance patient care and ultimately improve survival rates.

Assessing the Impact of a Public Health Intervention on Cancer Patients: A Trial Design and Sample Size Analysis

MATH5092M: Mixed Models with Medical Applications Report

Shrichand Bhuria
201800797

Department of Mathematics
University of Leeds
United Kingdom
April 30, 2024

1 Trial Design

In this section, we will outline the primary endpoint of the trial and the proposed design to achieve this objective. We will specify the underlying outcome model and conduct the primary analysis based on this model. Additionally, we will identify any assumptions made within this model and discuss the consequences of their violations. Finally, we will justify the chosen trial design and explain why it was preferred over alternative designs.

1.1 Primary Endpoint

The main objective of this study is to assess whether the implementation of the new public health intervention results in a later age at which patients receive their initial diagnosis, compared to a scenario with no intervention.

1.2 Trial Design

To evaluate the efficacy of the intervention in delaying the age at diagnosis for cancer patients, a clustered randomized trial design will be employed, with clustering based on the 12 hospital sites included in the 'CWSample' dataset. This dataset comprises personal information on 490 patients treated across these 12 hospital sites, along with details regarding their cancer characteristics. Information recorded includes the patient's ID numbers, the hospital where they were treated, the patient's age at the time of diagnosis, and the year of their initial diagnosis, typically ranging from approximately 30 to 80 years old.

Since the intervention operates at the hospital sites level, its scope extends beyond individuals at high risk of death to encompass all patients within those hospital sites. Consequently, data will be gathered on cancer patients receiving their initial diagnosis at each site as cases arise, along with the age at which diagnosis occurs. With each patient within a cluster assigned to the same trial arm, these clusters are considered nested within the intervention, constituting a nested design. It is assumed that the trial design maintains balance, with each arm containing an equal number of clusters, resulting in 6 clusters per arm and an equal number of patients denoted as n. After collecting all data, separate linear models will be fitted to the age of diagnosis for patients in the non-intervention clusters and those in the intervention clusters. These models, akin to Equation (3), will facilitate a comparison of estimated population means, with β_0 representing the average age of diagnosis in both trial arms. This comparison will reveal whether the age of diagnosis is higher in the intervention group compared to the non-intervention group.

The clustered randomized trial design entails implementing the intervention either across all patients in one hospital site or none within that site. This process involves pairing the six sites based on their respective average ages at diagnosis, as depicted in Table 1. The sites are ranked according to patient average age at diagnosis, with the highest-ranking pairs being selected first, followed by subsequent pairs. Thus, the pairs of SiteID are determined, such as 18 and 15, 20 and 16, 4 and 5, 8 and 19, 31 and 25, 10 and 7. Subsequently, the pairs are inputted into a random choice generator to select which site will receive the intervention. For instance, for the initial pair of 18 and 15, both SiteID names are entered into an online random choice generator wheel, and the site generated first will undergo intervention, while the other site will serve as the control.

SiteID	**4**	**5**	**7**	**8**	**10**	**15**	**16**	**18**	**19**	**20**	**25**	**31**
Average Age at Diagnosis (years)	45.5	45.4	41.2	43.7	42.3	47.2	46.2	49.3	43.4	46.7	42.8	42.9

Table 1: The patients average age at diagnosis in each hospital site.

One main reason for opting for a clustered randomized trial instead of an individually randomized one is that it's easier to either give the intervention to all patients at a hospital or not give it at all. If we split each hospital into two groups for the intervention, it could cause problems. For example, patients who didn't get the intervention might be influenced by those who did, or vice versa. This could change how patients behave, especially if those who got the intervention try to persuade others who didn't. This might affect the accuracy of the data we collect. Also, we want to see how the intervention affects patient outcomes across the whole hospital, which is simpler when everyone at a hospital gets the same treatment.

An alternative approach would be to use the existing data in the 'cancer.Rdata' dataset as the control group and implement the intervention across all hospital sites, recording patient deaths. While this method would decrease the amount of data needed, saving time and money, it would result in data from different time periods for patient age at diagnosis. Gathering new data from the intervention group would take many years, leading to potential discrepancies in comparing patient outcomes over time. Various factors, like advancements in healthcare leading to longer lifespans, could impact the average age at death differently across these time frames. Therefore, it's challenging to accurately compare data from different time periods, making it preferable to collect data for both control and intervention groups simultaneously.

1.3 Outcome Model

Assuming an equal number of clusters in each arm ($m = 6$) and an equal number of participants (n) in each arm, the outcome model implied by this trial design is described in equation (1). Here, the response variable y_{ijk} represents the age at death of the i-th patient in the j-th cluster in the k-th arm of the trial, where $i = 1, ..., p$, $j = 1, ..., m$, and $k = 0, 1$. This means there are $n = 6p$ participants in each arm, drawn from the p patients in each of the 6 hospital sites in each arm.

$$y_{ijk} \sim N(\mu_k + \mu_j, \sigma_e^2) \quad \text{with} \quad u_j \sim N(0, \sigma_u^2). \tag{1}$$

Expressing this model as a random intercept model without additional covariates yields:

$$y_{ijk} = \beta_0 + \beta_1 x_k + u_j + e_{ijk} \quad \text{with} \quad u_j \sim N(0, \sigma_u^2) \quad \& \quad e_{ijk} \sim N(0, \sigma_e^2), \tag{2}$$

where x_k is a binary indicator of the intervention arm ($x_k = 1$ for intervention, 0 otherwise).

2 Determining Sample Size Requirements

This section aims to establish the necessary parameter estimations for conducting sample size calculations. We will demonstrate these calculations using relevant formulas and R programming. Additionally, we will assess how variations in type I and type II error rates, treatment effects, and variance assumptions influence the required sample size.

2.1 Parameter Estimations

We will utilize the data from the 'CWsample' dataset to estimate the variance of the outcome model implied by the chosen trial design, along with the variance of patients' hospital SiteID. Assuming that patients' age at diagnosis follows a normal distribution, we determine the variance, σ^2, using a basic linear model:

$$y_i = \beta_0 + e_i \quad \text{with} \quad e_i \sim N(0, \sigma^2), \tag{3}$$

where the estimated population mean is $\beta_0 \approx 44.66$ years, representing the average patient's age at diagnosis. Moreover, the estimated variance is approximately $\sigma^2 \approx 7.37^2$. Additionally, assuming u_j follows a normal distribution with variance σ_u^2, as stated in equations (1) and (2), we determine this variance by creating a linear model for the patients' hospital sites of treatment:

$$u_j = \beta_0 + e_j \quad \text{with} \quad e_j \sim N(0, \sigma_u^2), \tag{4}$$

where β_0 represents the estimated population mean. Implementing this model in R provides an estimated variance of $\sigma_u^2 \approx 3.50^2$, which will be utilized in the sensitivity analysis to compute the intracluster correlation coefficient.

2.2 Determining Sample Size

We initiate by computing the sample size necessary for type I and II error rates of 0.025 (one-sided) and 0.2, correspondingly. Following this, we will explore alternative nominal error rates and their impact on the required sample size. The formula employed for calculating the sample size is as follows:

$$n = \left(\frac{\sqrt{2\sigma^2}(z_{1-\alpha} - z_\beta)}{\delta_1 - \delta_0} \right)^2 \tag{5}$$

Given our objective to increase the patients' expected age at death by 2.5 years, we formulate the null hypothesis $H_0 : \delta = \delta_0$ and the alternative hypothesis $H_1 : \delta = \delta_1$. Here, δ represents the treatment effect, calculated as $\delta = \mu_0 - \mu_1$, signifying the difference in the average age at diagnosis before and after intervention. Since our aim is to raise the age at diagnosis by 2.5 years, we set the original treatment effect before intervention as $\delta_0 = 0$, and the proposed treatment effect as $\delta_1 = 2.5$. Substituting the type I and type II errors, $\alpha = 0.025$ (one-sided) and $\beta = 0.2$ respectively, along with the estimated variance σ^2, and the values for δ_0 and δ_1, we compute the sample size:

$$n = \left(\frac{\sqrt{2(9.62)^2}(z_{1-0.025} - z_{0.2})}{2.5 - 0} \right)^2 = 136.46 \tag{6}$$

where, $z_x = \Phi^{-1}(x)$ represents the inverse standard normal distribution. Here, $z_{0.975} \approx -1.96$ and $z_{0.2} \approx 0.842$. Since this value represents the minimum number of patients needed, we round it to obtain $n = 137$ participants for the specified type I and type II errors. This calculation can also be performed using the 'power.t.test' function in R. By specifying $\delta = 2.5$, standard deviation $\sigma = 7.37$, and power $1 - \beta = 0.8$, we find the required sample size to be $n = 137.39$, hence we need $n = 138$ patients according to this function. The slight difference in the number of patients required in each arm can be attributed to the estimations of the inverse standard normal distribution used, rather than the exact values utilized within the R function.

From Fig. 1, it's evident that a higher type I error rate corresponds to a smaller required sample size. Conversely, a higher power value leads to a larger required sample size, while a lower type II error rate results in a higher number of patients needed. Additionally, considering the age of first death as a normally distributed variable based on the 'CWsample' dataset, with a variance of approximately $\sigma \approx 7.37^2$, we can examine how altering this variance affects the required sample size, as depicted in Fig. 2 below.

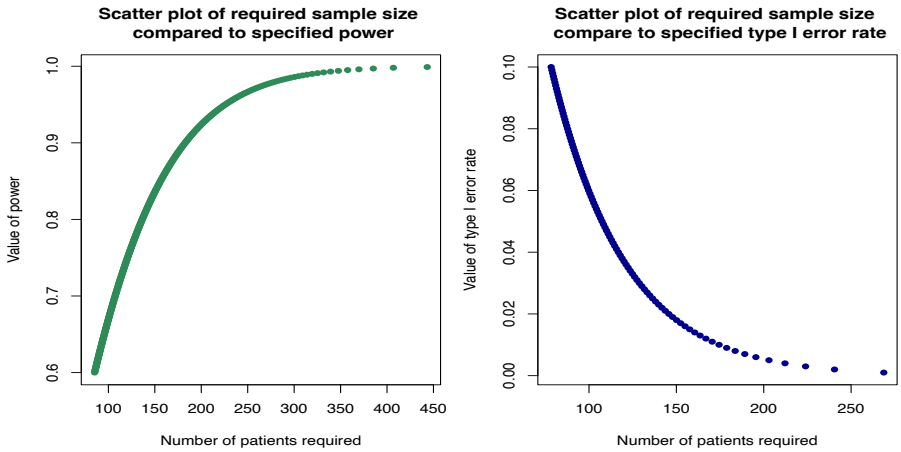


Figure 1: Scatter plot illustrating the relationship between required sample size and the specified power value, $1-\beta$ (left), and the specified type I error rate (right).

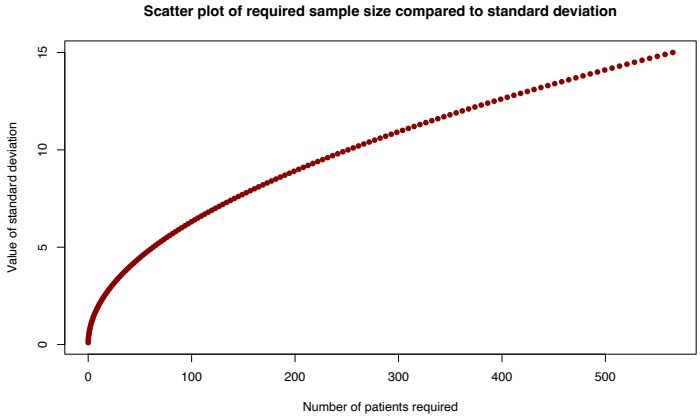


Figure 2: Scatter plot showing the relationship between required sample size and the assumed standard deviation.

2.3 Sensitivity Analysis

After determining the variance of age at diagnosis as $\sigma^2 \approx 7.37^2$ and the variance of SiteID as $\sigma_u^2 \approx 3.5^2$, we can calculate the residual variance as $\sigma_e^2 \approx 7.37^2 - 3.5^2 \approx 6.68^2$, since the total variance is the sum of the individual variances. Using this model, we estimate the intracluster correlation coefficient to be:

$$\rho = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2} \approx \frac{3.50^2}{7.37^2} \approx 0.225. \tag{7}$$

Considering clusters with this intracluster correlation coefficient of $\rho \approx 0.225$, we calculate a design effect of $1 - \rho \approx 0.775$. Multiplying this by our proposed sample size of $n = 136.46$, we find the required sample size decreases to $n = 106$ per arm. However, since these calculations are based on assumptions regarding the normal distribution of age at diagnosis and state variables, and the variances are only estimates, reducing the sample size to this extent may pose risks. Therefore, we will maintain the original required sample size of $n = 137$ in each arm to ensure the accuracy of our results. With a sample size of 137 for each arm of the trial, we can confidently proceed with evaluating the effectiveness of the public health intervention on cancer patients. This sample size provides us with sufficient statistical power to detect meaningful differences in the age at diagnosis between the intervention group and the control group. Additionally, by adhering to this sample size, we enhance the reliability and accuracy of our findings, ensuring robust conclusions regarding the intervention's impact on cancer patient outcomes.