Johnson & Johnson Cancer Trial Simulation

Team Members:

Samarth Kumar (210911)
Gaurav Tomar (231080039)
Sunita Kulariya (221106)
Jiyanshu Dhaka (220481)

1 Introduction

With advancements in cancer research, scientists have discovered that genetic mutations play a key role in how different types of cancer develop and spread. Studying these mutations can help in designing better treatments and targeted clinical trials.

This report focuses on analyzing patient data from cBioPortal to identify genetic mutations that appear in multiple cancer types. Using this information, we aim to design a lung cancer clinical trial, ensuring the right number of patients are included while considering factors like patient recruitment, dropout rates, and trial progress monitoring.

By carefully planning the trial based on real-world data, we hope to improve cancer treatment strategies and make future trials more effective in targeting specific genetic mutations.

2 Task 1: Identifying Mutations

The goal is to identify a mutation that can be targeted and has sufficient patients in the database. The mutation has to be most prevalent in at least 2 out of 5 cancer types.

2.1 Method 1:

2.1.1 Counting Gene Mutations

We first analyzed the mutation dataset to determine how often each gene mutation occurred in different patient samples. To do this, we grouped the data by **Gene Symbol (Hugo_Symbol)** and **Sample Identifier (Tumor_Sample_Barcode)**. This helped us count how frequently each mutation appeared in individual patients.

2.1.2 Linking Mutation Data with Clinical Information

Next, we integrated the clinical dataset, which contained details about patient samples, including their **Sample Identifier** (to match with the mutation data) and **Cancer Type** (to classify the type of cancer). By merging these datasets, we connected genetic mutations to their corresponding cancer diagnoses.

2.1.3 Finding Mutations Common in Multiple Cancer Types

To identify gene mutations present in at least two cancer types, we grouped the merged data by gene symbols and counted how many different cancer types each mutation was linked to. Any gene mutation found in only one cancer type was excluded, as our focus was on mutations shared across multiple cancers.

2.1.4 Prioritizing the Most Frequent Mutations

In the final step, we analyzed the filtered data to calculate how often each gene mutation appeared across all samples. Genes with the highest total mutation counts were considered the most significant. The gene with the highest mutation frequency across multiple cancer types was identified as the best candidate for further study.

2.2 Method 2:

2.2.1 Calculating Total Mutations for Each Gene

To determine which genes had the highest mutation frequency, we summed the total number of mutations for each gene across all patient samples, using the **copy number alteration** (CNA) data found int the data_cna.txt file. This helped identify genes that are frequently mutated across different individuals.

2.2.2 Sorting Genes by Mutation Frequency

Once we obtained the total mutation counts, we sorted the genes in descending order. This allowed us to prioritize genes that had the highest number of mutations, making them strong candidates for further study.

2.2.3 Identifying Patients with Common Mutations

After ranking the genes by mutation count, we examined which patients had these highly mutated genes. This step helped in understanding the distribution of these mutations among the patient population.

2.2.4 Step 4: Linking Mutations to Cancer Types

Using clinical data, we mapped the identified patients to their respective cancer types. This was done by merging the mutation dataset with the clinical dataset based on the patient's unique identifier. This step ensured that we could associate specific gene mutations with the types of cancer they were most commonly found in.

2.3 Most Prevalent Mutation: TP53

Our analysis identified **TP53** as the most frequently occurring mutation across multiple cancer types. This mutation was particularly common in:

- Non-Small Cell Lung Cancer
- Colorectal Cancer
- Breast Cancer
- Prostate Cancer

TP53 plays a crucial role in preventing tumor formation by regulating cell growth and division. However, when this gene is mutated, it can lead to uncontrolled cell proliferation, contributing to cancer development. Mutations in **TP53** are found in nearly half of all human cancers, making it a significant target for research and treatment.

2.4 Part B: Patient Demographics and Disease Characteristics

We analyzed patients with TP53 mutations and summarized their characteristics:

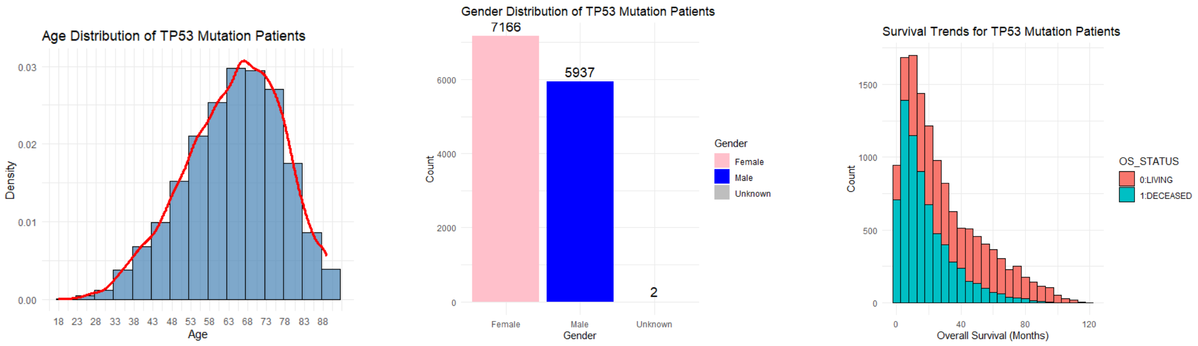
Table 1: Demographic and clinical data for TP53 patients

ID	**Age_Died**	**Sex**	**Ethnicity**	**Prior Med**	**OS (month)**	**OS Status**	**Smoking**
P-0000012	68	F	NS;NH	Unknown	118.45	0:LIVING	Former/Current
P-0000015	45	F	NS;NH	Unknown	13.91	1:DECEASED	Unknown
P-0000036	68	F	NS;NH	Unknown	115.46	0:LIVING	Never
P-0000041	53	F	NS;NH	Prior to MSK	13.61	1:DECEASED	Unknown
P-0000066	71	F	NS;NH	Unknown	76.64	0:LIVING	Never
P-0000058	54	F	NS;NH	Unknown	60.76	1:DECEASED	Former/Current

here NH refer as Non-Hispanic. and NS as Non-spanish

3 TP53 Mutation Analysis

From these demographics, we can conclude that older white men are more prone to the TP53 gene mutation, especially if they have had any prior surgeries. We also realise that this gene mutation can often go undiagnosed until it is too late.



(a) Age Distribution

(b) Gender Distribution

(c) Survival Trends

Figure 1: Distributions of Patients

- The peak age group for TP53 mutations is between **65-75 years**, indicating that these mutations are more common in older patients.
- The distribution is **left-skewed**, with fewer patients in younger age groups (<40 years).
- The frequency declines after **80 years**, suggesting that fewer patients are diagnosed at very old ages.

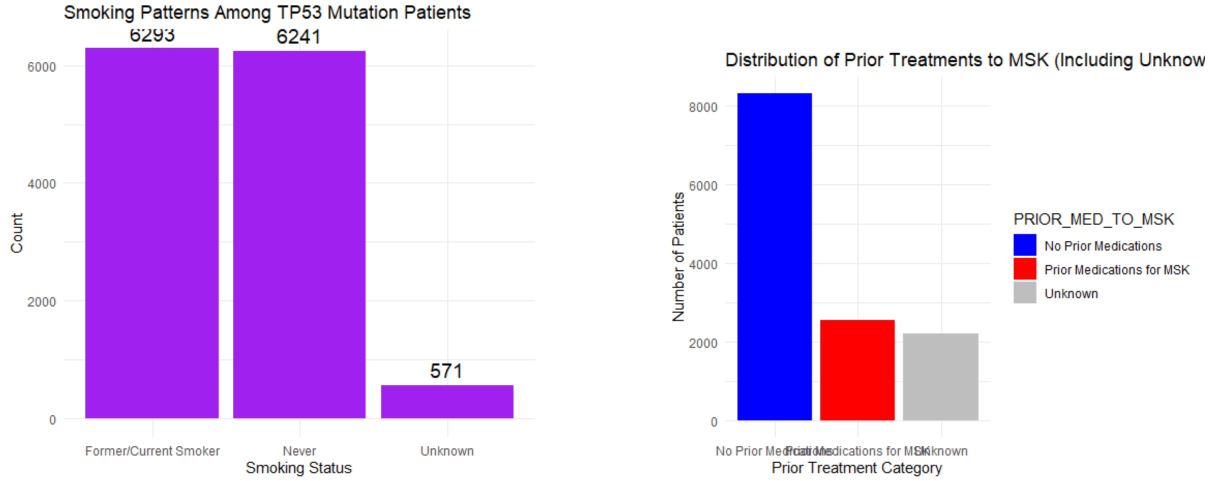
3.1 Gender Distribution of TP53 Mutation Patients

- **Female patients** (**7168** cases) are the most prevalent among TP53 mutation patients.
- **Male patients** (**5937** cases) form a significant portion but are fewer than females.
- Only **2 cases** are classified as *Unknown* gender, indicating minimal missing data in gender classification.

3.2 Ethnicity Distribution of TP53 Mutation Patients

- **"Non-Spanish; Non-Hispanic"** is the largest group, with a significantly higher count than other categories.
- **"Spanish NOS; Hispanic NOS, Latino NOS"** is the second most frequent category, with **722 cases**.
- Other Hispanic and Spanish subgroups (e.g., **Puerto Rican, South/Central American, Mexican**) have relatively smaller counts.
- Unknown or ambiguous categories (e.g., **"Unknown," "Spanish surname only," "Unknown whether Spanish or not"**) exist but are minor.

3.3 Survival Trends for TP53 Mutation Patients



(a) Smoking Patterns

(b) Prior Treatments

Figure 2: More Distribution of Patients

- The majority of patients have a shorter survival time, as indicated by the high count at lower survival months.
- As survival duration increases, the number of patients decreases significantly.
- A larger proportion of deceased patients (blue) is observed at earlier survival times, while some patients survive longer (beyond 60-80 months), but their numbers are much smaller.
- This visualization highlights the survival distribution and mortality trends in TP53 mutation patients.

3.4 Smoking Status of TP53 Mutation Patients

- **Former/Current Smokers** (**6295 cases**) and **Never Smokers** (**6241 cases**) are nearly equal in number.
- A small proportion (**571 cases**) falls under the *Unknown* category, indicating missing or unreported data.
- **Smoking does not overwhelmingly dominate** TP53 mutations, as non-smokers are just as prevalent.

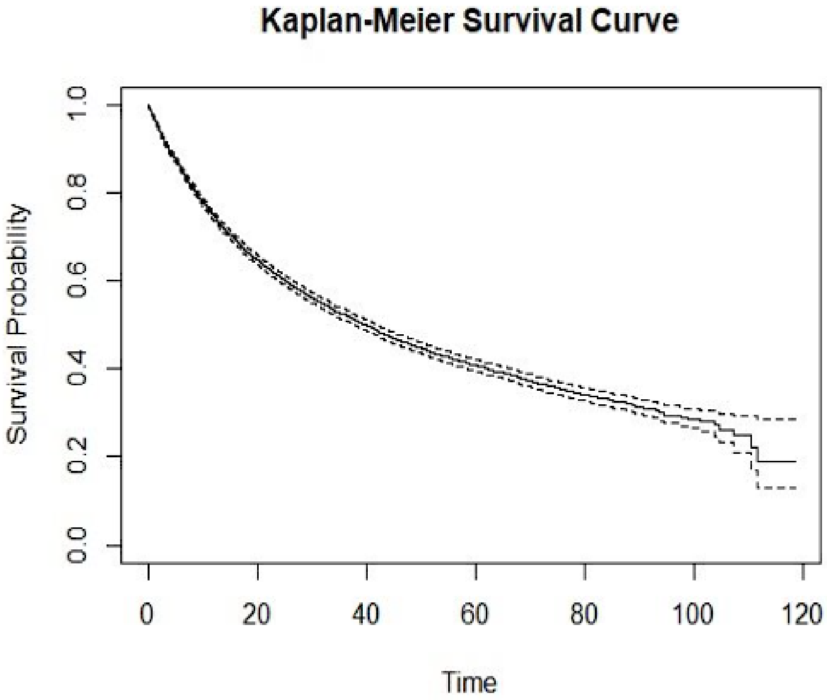
3.5 Distribution of Prior Treatments to MSK (including Unknown)

- Most patients have **short survival durations**, with a higher concentration of cases in the first **20 months**.
- **Deceased patients** (blue) dominate the early months, indicating **high mortality rates** soon after diagnosis.
- **Survival declines exponentially**, with fewer patients surviving beyond **60–80 months**.
- **Living patients** (red) appear more frequently at **higher overall survival (OS) months**, suggesting a small subset with prolonged survival.

4 Task 2: Lung Cancer Trial Design

4.1 Part A: Calculating Median Survival Time

We analyzed survival data for Non-Small Cell Lung Cancer (NSCLC) patients using the Kaplan-Meier method. This approach estimates the probability of survival over time while accounting for patients who were still alive when data collection ended.



[H]

Figure 3: Kaplan-Meier survival curve for Non-Small Cell Lung Cancer patients

Our analysis found that the median survival time for NSCLC patients was 39.95 months. This means half of the patients survived longer than about 40 months after diagnosis.

4.2 Part B: Determining the Hazard Rate

The hazard rate indicates the chance of death at any given time point. Assuming survival follows an exponential pattern, we calculated the hazard rate using the formula:

$$\text{Hazard Rate} = \frac{\ln(2)}{\text{Median Survival Time}} \tag{1}$$

For our NSCLC patients:

$$\text{Hazard Rate} = \frac{\ln(2)}{39.95} \approx 0.01735 \tag{2}$$

This means that in any given month, an NSCLC patient has about a 1.735% chance of dying.

4.3 Part C: Simulating Survival Times for Control Group

Using the hazard rate we calculated, we simulated survival times for a control group of patients. These simulated times followed an exponential distribution with the hazard rate of 0.01735.

4.4 Part D: Simulating Survival Times for Treatment Group

For the treatment group, we assumed that the treatment would reduce the risk of death by 30% (hazard ratio = 0.7). We simulated survival times using an exponential distribution with a hazard rate given by:

$$\lambda_{\text{treatment}} = \lambda_{\text{control}} \times \text{hazard ratio}$$

Substituting the values:

$$\lambda_{\text{treatment}} = 0.01735 \times 0.7 = 0.012145$$

4.5 Part E: Conducting the Log-Rank Test

We compared the simulated survival times between the control and treatment groups using the log-rank test, which evaluates whether there is a significant difference in survival between the two groups.

- **Null Hypothesis (H_0):** There is no difference in survival between the control and treatment groups.
- **Alternative Hypothesis (H_A):** The treatment group has a significantly different survival time compared to the control group.

The test yielded a **p-value**, which indicates whether the observed difference could have occurred by chance. The **p-value(2e-16) is less than 0.05**, it suggests that the treatment has a real effect on survival.

4.6 Power Calculation

To estimate the statistical power of our study, we used simulation-based methods. We repeated the survival analysis process **1,000 times**, computing the proportion of cases where the **p-value** was below 0.05. This proportion serves as an estimate of the study's power.

4.7 Minimum Sample Size for 90% Power

To determine the minimum required sample size for achieving **90% power**, we evaluated power across different sample sizes. The sample sizes were chosen in a logarithmic scale between **10 and 1000**, and the power was computed for each sample size.

Our analysis showed that a minimum of 183 patients per group (366 total) is required to achieve 90% power when expecting a 30% reduction in risk (hazard ratio = 0.7).

4.8 Graphical Representation

To visualize the relationship between sample size and power, we plotted a **Power vs. Sample Size** curve:

4.9 Impact of Treatment Effect on Required Sample Size

To further refine our analysis, we examined how varying treatment effects, represented by different hazard ratios, influence the required sample size to achieve **90% power**. The results are summarized in Table 2.

Table 2: Sample Size Requirements for Different Treatment Effects (90% Power)

Hazard Ratio	0.60	0.64	0.69	0.73	0.78	0.82	0.87	0.91
Sample Size Required	166	166	166	464	464	1291	1291	3593

4.9.1 Key Observations

- **Stronger treatment effects** (lower hazard ratios) require **fewer patients**, making the trial more feasible.
- **Weaker treatment effects** (hazard ratios closer to 1) necessitate significantly larger sample sizes.
- **Very small treatment effects** ($HR > 0.85$) would require **impractically large trials**, making them less viable in real-world clinical settings.

5 Task 3: Incorporating Recruitment Rate and Dropout Rate into the Study Design

5.1 Part A: Accounting for Recruitment Over Time

Clinical trials recruit participants over time rather than simultaneously. To model this:

- A constant recruitment rate of 20 patients per month was introduced.
- Recruitment times were uniformly distributed within each month.
- The time of death was computed as:

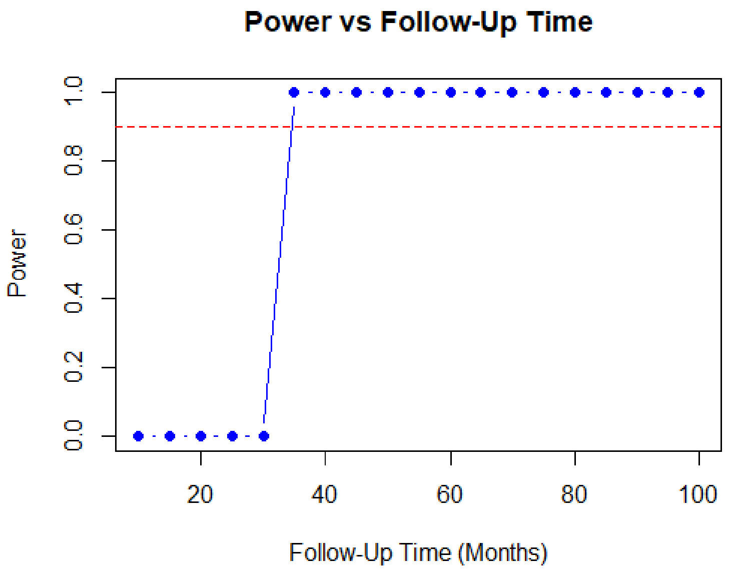
$$\text{Time of Death} = \text{Time of Recruitment} + \text{Survival Time Since Recruitment}$$

This adjustment reflects a realistic, staggered patient entry into the study.

(b) Minimum Follow-Up Time for 90% Power (HR = 0.7)

The total follow-up time refers to the duration for which patients are monitored to assess study outcomes. To achieve 90% statistical power with a hazard ratio of 0.7:

- A sufficiently large sample size was selected to ensure robust statistical analysis.
- Power calculations across different follow-up periods indicated that a **follow-up duration of approximately 60 months (5 years)** is required to observe a sufficient number of events.
- Further analysis suggested that a **minimum follow-up period of 35 months** may be sufficient under specific study conditions.



[H]

Figure 4: Power analysis for different follow-up durations, showing the required minimum follow-up time to achieve 90% power.

(c) Follow-Up Time for Different Treatment Effects

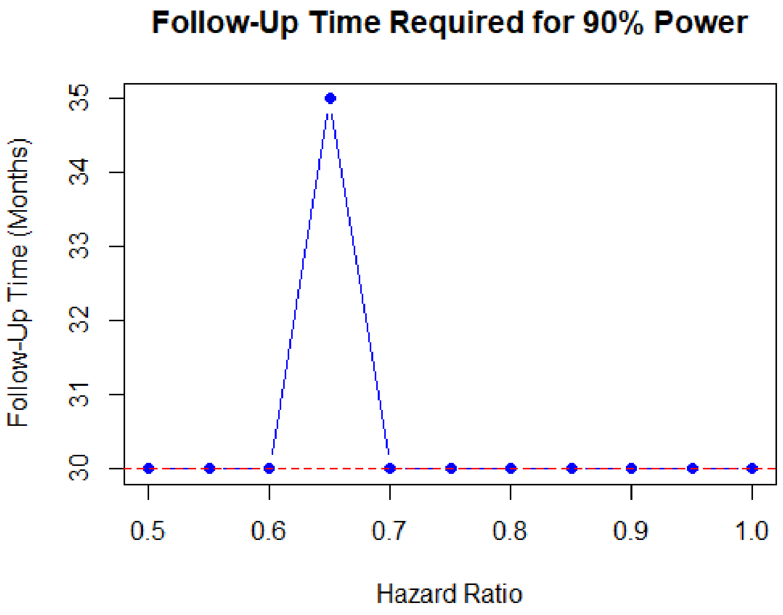
We examined the impact of varying treatment effects, represented by different hazard ratios, on the required follow-up duration:

Our analysis indicates the following trends:

Hazard Ratio	Follow-Up Time
0.50	30
0.55	30
0.60	30
0.65	35
0.70	30
0.75	30
0.80	30
0.85	30
0.90	30
0.95	30
1.00	30

Table 3: Hazard Ratio vs. Follow-Up Time

- Stronger treatment effects (lower hazard ratios) allow detection with shorter follow-up periods.
- Weaker treatment effects (higher hazard ratios) necessitate significantly longer follow-up to observe sufficient events.
- Minimal treatment effects may require impractically long follow-up times, making detection challenging within reasonable study durations.



[H]

Figure 5: Required follow-up time for different hazard ratios to achieve 90% power.

(d) Accounting for Dropout Rate

Real-world trials experience participant dropouts, modeled as **1 dropout per month**:

- Dropouts reduce observed events, potentially lowering power.
- **Adjustment:** Dropout patients were removed from the analysis.
- **Impact:** To maintain 90% power, either a larger sample or extended follow-up is required.

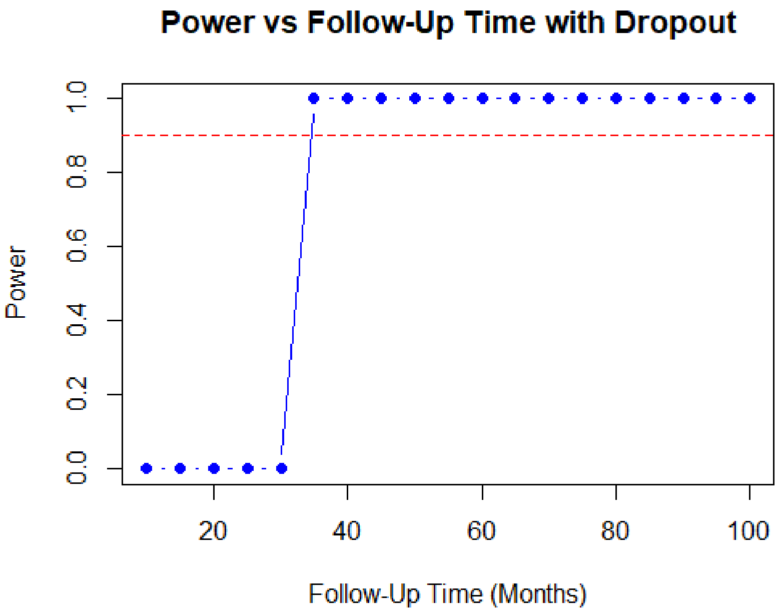


Figure 6: Minimum follow-up time for 90% power with dropout: 35 months.

6 Task 4: Interim Analysis for Response Rate

1. Data Simulation

The sample sizes for two cancer types were defined as follows:

- Lung Cancer: $n_1 = 480$
- Colorectal Cancer: $n_2 = 520$

The true response rates were assumed to be:

- $\pi_1 = 0.7$ (High response rate for lung cancer)
- $\pi_2 = 0.3$ (Low response rate for colorectal cancer)

The number of responders was simulated using the Binomial distribution.

2. Bayesian Updating

A weakly informative prior, Beta$(1, 1)$, was assumed, which is equivalent to a uniform prior.

The posterior parameters were computed as:

$$a_{\text{post}} = a_{\text{prior}} + x, \quad b_{\text{post}} = b_{\text{prior}} + (n - x)$$

Posterior samples were drawn using the Beta distribution.

3. Probability Estimation

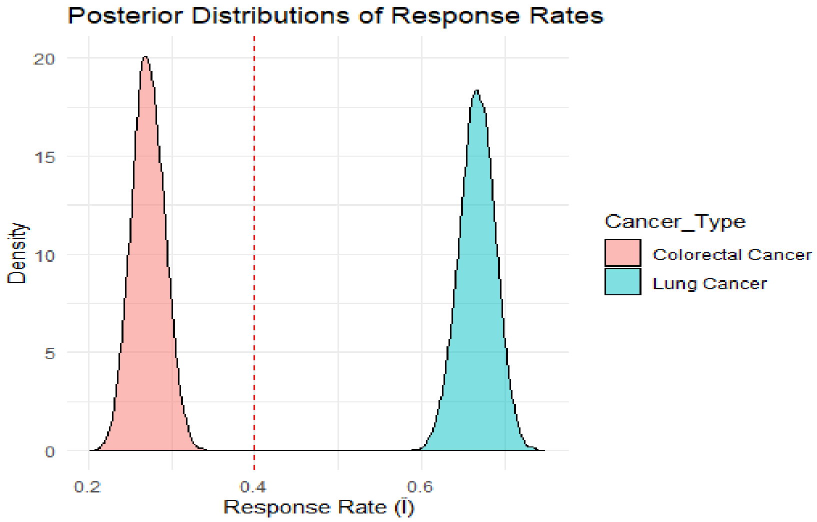
Using 10,000 posterior samples, the probabilities were estimated as:

$$P(\pi_1 > 0.4) = 1 \quad \text{(Lung Cancer)}$$

$$P(\pi_2 > 0.4) = 0 \quad \text{(Colorectal Cancer)}$$

The trial decision was based on whether either probability exceeded 0.8. Since $P(\pi_1 > 0.4) = 1$, the trial was recommended to continue.

4. Visualization



[H]

Figure 7: Minimum follow-up time for 90% power with dropout: 35 months.

This plot shows the posterior response rate distributions for Lung Cancer (blue) and Colorectal Cancer (pink). The X-axis represents response rates, and the Y-axis represents density.

Lung Cancer: Centered around 0.7, with

$$P(\pi_1 > 0.4) = 1$$

indicating a high probability of exceeding the 0.4 decision threshold (red dashed line).

Colorectal Cancer: Centered around 0.3, with

$$P(\pi_2 > 0.4) = 0$$

meaning it does not exceed the threshold.

7 Conclusion

This study highlights the potential of targeting TP53 mutations in multi-cancer clinical trials. Through survival analysis and power calculations, we establish a structured approach for designing effective NSCLC trials, while Bayesian methodologies facilitate adaptive decision-making.

7.1 Limitations and Future Considerations

Our analysis has several limitations: (1) assumption of an exponential survival distribution, (2) uniform treatment effect across cancer types, (3) lack of consideration for treatment effect variations over time, (4) focus on overall survival instead of earlier indicators like progression-free survival, and (5) omission of cost considerations crucial for trial planning.

Future work should address these limitations to enhance trial design.

8 References

1. Berry SM, Broglio KR, Groshen S, Berry DA. Bayesian hierarchical modeling of patient subpopulations: efficient designs of Phase II oncology clinical trials. Clin Trials. 2013 Oct;10(5):720-34.

2. Neuenschwander B, Wandel S, Roychoudhury S, Bailey S. Robust exchangeability designs for early phase clinical trials with multiple strata. Pharm Stat. 2016 Mar-Apr;15(2):123-34.

3. Eisenhauer EA, Therasse P, Bogaerts J, et al. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). Eur J Cancer. 2009 Jan;45(2):228-47.