Analisi Completa del Dataset di Tumori Cerebrali Pediatrici

Data Cleaning e Data Exploration con Analisi Statistica

Nicholas Mondella 859673

2025-09-16

Contents

1 Introduzione e Obiettivi dello Studio **4**

1.1 Contesto Clinico 4

1.2 Obiettivi Primari 4

1.3 Metodologia Statistica 4

2 Descrizione del Dataset e delle Variabili **5**

2.1 Caratteristiche Generali del Dataset 5

2.2 Classificazione delle Variabili Cliniche 5

2.2.1 Variabili Demografiche 5

2.2.2 Variabili di Presentazione Clinica 5

2.2.3 Variabili Anatomopatologiche 5

2.2.4 Variabili di Trattamento 5

2.2.5 Variabili di Outcome 6

3 Analisi della Qualità dei Dati **7**

3.1 Valutazione Complessiva dei Valori Mancanti 7

3.1.1 Interpretazione dei Valori Mancanti 8

3.2 Visualizzazione Pattern dei Valori Mancanti 8

3.2.1 Spiegazione dei Valori Mancanti Clinicamente Rilevanti 9

4 Data Cleaning: Strategia di Trattamento **10**

4.1 Strategia Clinicamente Informata 10

4.2 Applicazione delle Strategie di Cleaning 11

4.2.1 Rimozione Variabili con Eccessivi Valori Mancanti 11

4.2.2 Imputazione per Variabili Numeriche 11
4.2.3 Gestione Outliers con Approccio Clinico 12

5 Analisi Esplorativa: Caratteristiche Demografiche 14
5.1 Distribuzione per Sesso 14
5.1.1 Interpretazione Clinica della Distribuzione per Sesso 15
5.2 Analisi dell'Età dei Pazienti 15
5.2.1 Significato Clinico dell'Analisi dell'Età 17

6 Analisi della Presentazione Clinica 18
6.1 Durata dei Sintomi Prima della Diagnosi 18
6.1.1 Interpretazione Clinica della Durata dei Sintomi 20
6.2 Analisi dei Sintomi Neurologici 20
6.2.1 Significato Clinico dei Sintomi Neurologici 21

7 Analisi degli Istotipi Tumorali 22
7.1 Distribuzione degli Istotipi 22
7.1.1 Significato Clinico degli Istotipi 23
7.2 Correlazione tra Istotipo e Localizzazione 24
7.2.1 Implicazioni Cliniche della Localizzazione 24

8 Analisi delle Strategie Terapeutiche 25
8.1 Approcci Chirurgici 25
8.2 Radioterapia 26
8.2.1 Significato Clinico delle Dosi di Radioterapia 27
8.3 Chemioterapia 27
8.3.1 Razionale Clinico dei Diversi Timing di Chemioterapia 28

9 Analisi di Sopravvivenza 29
9.1 Overall Survival (OS) 29
9.1.1 Interpretazione Clinica dei Dati di Sopravvivenza 31

10 Analisi delle Correlazioni Statisticamente Significative 32
10.1 Matrice di Correlazione delle Variabili Numeriche 32
10.2 Test di Significatività delle Correlazioni ($p < 0.005$) 34
10.2.1 Interpretazione Clinica delle Correlazioni Significative 38

11 Test Statistici per Differenze tra Gruppi **39**

11.1 Confronto tra Istotipi e Outcome Clinici 39

11.1.1 Test per Sopravvivenza tra Gliomi di Alto e Basso Grado 39

11.1.2 Significato Clinico del Confronto LGG vs HGG 40

11.2 Analisi della Sopravvivenza per Localizzazione 40

11.3 ANOVA per Confronto Multiple di Istotipi 41

11.3.1 Interpretazione Clinica dell'ANOVA 42

12 Modello Predittivo di Sopravvivenza **43**

12.1 Regressione Lineare Multipla per Predizione OS 43

12.2 Diagnostica del Modello 45

12.2.1 Interpretazione dei Coefficienti del Modello 45

13 Salvataggio del Dataset Pulito **46**

13.1 Esportazione dei Dati Processati 46

14 Conclusioni e Considerazioni Cliniche **48**

14.1 Principali Scoperte dell'Analisi 48

14.1.1 Caratteristiche Demografiche e Cliniche 48

14.1.2 Pattern di Presentazione Clinica 48

14.1.3 Distribuzione degli Istotipi 48

14.1.4 Outcome di Sopravvivenza 48

14.2 Implicazioni per la Pratica Clinica 48

14.2.1 Stratificazione Prognostica 48

14.2.2 Ottimizzazione Terapeutica 48

14.2.3 Follow-up e Sorveglianza 49

14.3 Limitazioni dello Studio 49

14.3.1 Limitazioni Metodologiche 49

14.3.2 Limitazioni Statistiche 49

14.3.3 Generalizzabilità 49

14.4 Raccomandazioni per Ricerche Future 49

14.4.1 Studi Prospettici 49

14.4.2 Analisi Avanzate 49

14.4.3 Studi Multicentrici 50

14.5 Bibliografia 50

1 Introduzione e Obiettivi dello Studio

1.1 Contesto Clinico

Questo studio presenta un'analisi approfondita di un dataset clinico contenente informazioni su **174 pazienti pediatrici** affetti da tumori cerebrali. I tumori cerebrali rappresentano la seconda forma più comune di neoplasia in età pediatrica e costituiscono la principale causa di morte per cancro nei bambini.

1.2 Obiettivi Primari

1. **Data Cleaning**: Identificare e correggere problematiche nella qualità dei dati
2. **Data Exploration**: Analizzare le caratteristiche demografiche e cliniche dei pazienti
3. **Analisi Statistica**: Identificare relazioni significative tra variabili ($p < 0.005$)
4. **Insights Clinici**: Fornire evidenze per supportare decisioni cliniche

1.3 Metodologia Statistica

L'analisi adotta una **soglia di significatività rigorosa ($p < 0.005$)** per minimizzare il rischio di falsi positivi, seguendo le raccomandazioni più recenti per la ricerca biomedica.

2 Descrizione del Dataset e delle Variabili

2.1 Caratteristiche Generali del Dataset

Il dataset comprende **34 variabili** che descrivono diversi aspetti del percorso clinico dei pazienti, dalla presentazione iniziale agli outcome di sopravvivenza.

2.2 Classificazione delle Variabili Cliniche

Le variabili del dataset possono essere classificate in diverse categorie cliniche:

2.2.1 Variabili Demografiche

- **Pat. No.**: Identificativo univoco del paziente (*utilizzato solo come identificativo, escluso dalle analisi statistiche*)
- **Sex**: Sesso del paziente (1 = Maschio, 2 = Femmina)
- **Age**: Anno di nascita del paziente

2.2.2 Variabili di Presentazione Clinica

- **Duration of symptoms before diagnosis**: Durata dei sintomi prima della diagnosi (giorni)
- **Increased ICP**: Presenza di ipertensione intracranica (1 = Sì, 2 = No)
- **Epileptic seizures**: Presenza di crisi epilettiche (1 = Sì, 2 = No)
- **Neurological deficit**: Presenza di deficit neurologici (1 = Sì, 2 = No)
- **Hormonal abnormalities**: Presenza di alterazioni ormonali (1 = Sì, 2 = No)

2.2.3 Variabili Anatomopatologiche

- **Embryonal tumors**: Tumori embrionali (1 = Sì, 2 = No)
- **HGG**: Gliomi di alto grado (1 = Sì, 2 = No)
- **LGG**: Gliomi di basso grado (1 = Sì, 2 = No)
- **Craniopharyngeoma**: Craniofaringioma (1 = Sì, 2 = No)
- **GCT / NGCT**: Tumori a cellule germinali (1 = Sì, 2 = No)
- **Ependymoma**: Ependimoma (1 = Sì, 2 = No)

2.2.4 Variabili di Trattamento

- **Operation type**: Tipo di intervento chirurgico
- **Radiotherapy**: Radioterapia (1 = Sì, 2 = No)
- **Neoadjuvant HT**: Chemioterapia neoadiuvante (1 = Sì, 2 = No)

- **Concomitant HT**: Chemioterapia concomitante (1 = Sì, 2 = No)
- **Adjuvant HT**: Chemioterapia adiuvante (1 = Sì, 2 = No)

2.2.5 Variabili di Outcome

- **OS**: Overall Survival (sopravvivenza globale in mesi)
- **Status OS**: Stato di sopravvivenza (1 = Deceduto, 2 = Vivo)

3 Analisi della Qualità dei Dati

3.1 Valutazione Complessiva dei Valori Mancanti

Prima di procedere con qualsiasi analisi statistica, è fondamentale valutare la **completezza del dataset** e identificare i pattern dei valori mancanti.

```
missing_summary <- data.frame(
  Variabile = names(raw_data),
  Valori_Mancanti = sapply(raw_data, function(x) sum(is.na(x))),
  Percentuale = round(sapply(raw_data, function(x) sum(is.na(x)) / length(x) * 100), 2),
  stringsAsFactors = FALSE
)
missing_summary <- missing_summary[order(missing_summary$Percentuale, decreasing = TRUE), ]

knitr::kable(missing_summary[missing_summary$Valori_Mancanti > 0, ],
             caption = "Variabili con Valori Mancanti nel Dataset",
             booktabs = TRUE, row.names = FALSE)
```

Table 1: Variabili con Valori Mancanti nel Dataset

Variabile	Valori_Mancanti	Percentuale
CSF cytology finding	99	56.90
Boost dose	88	50.57
CSI dose	87	50.00
Operation type	21	12.07
Total dose	10	5.75
MRI finding	4	2.30
Concomitant HT	4	2.30
Pat. No.	1	0.57
Sex	1	0.57
Age	1	0.57
Duration of sympthoms before diagnosis (days)	1	0.57
Increased ICP	1	0.57
Epileptic seizures	1	0.57
Neurological deficit	1	0.57
Hormonal abnormalities	1	0.57
Localisation	1	0.57
CSF cytology	1	0.57
Extent of disease	1	0.57

Variabile	Valori_Mancanti	Percentuale
Neoadjuvant HT	1	0.57
Radiotherapy	1	0.57
Adjuvant HT	1	0.57
OS	1	0.57
Status OS	1	0.57

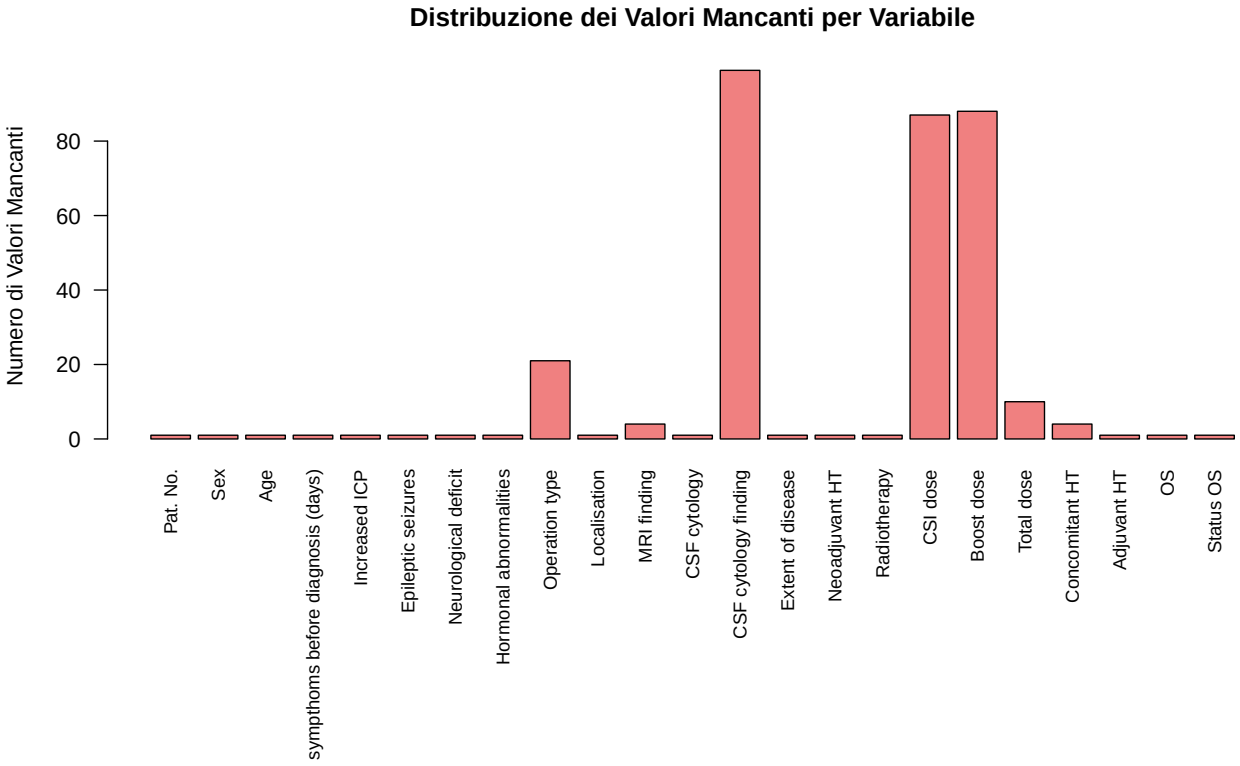
3.1.1 Interpretazione dei Valori Mancanti

I valori mancanti nel contesto clinico possono avere diversi significati:

1. **Missing Completely at Random (MCAR)**: Dati persi casualmente
2. **Missing at Random (MAR)**: Dati mancanti dipendenti da altre variabili osservate
3. **Missing Not at Random (MNAR)**: Dati mancanti correlati al valore stesso

3.2 Visualizzazione Pattern dei Valori Mancanti

```
missing_counts <- sapply(raw_data, function(x) sum(is.na(x)))
vars_with_missing <- missing_counts[missing_counts > 0]

if(length(vars_with_missing) > 0) {
  par(mar = c(12, 4, 4, 2))
  barplot(vars_with_missing,
          main = "Distribuzione dei Valori Mancanti per Variabile",
          ylab = "Numero di Valori Mancanti",
          las = 2, col = "lightcoral",
          cex.names = 0.8)
  par(mar = c(5, 4, 4, 2))
}
```



3.2.1 Spiegazione dei Valori Mancanti Clinicamente Rilevanti

- **CSF cytology finding**: Mancante quando l'esame del liquor non è stato eseguito
- **CSI dose/Boost dose**: Mancante quando non è stata somministrata radioterapia
- **Operation type**: Mancante quando non è stato eseguito intervento chirurgico

4 Data Cleaning: Strategia di Trattamento

4.1 Strategia Clinicamente Informata

La strategia di cleaning deve rispettare il **significato clinico** dei valori mancanti:

```
cleaned_data <- raw_data

cat("=== STRATEGIA DI TRATTAMENTO VALORI MANCANTI ===\n")
```

```
## === STRATEGIA DI TRATTAMENTO VALORI MANCANTI ===
```

```
cat("Totale pazienti nel dataset:", nrow(cleaned_data), "\n\n")
```

```
## Totale pazienti nel dataset: 174
```

```
for(col in names(cleaned_data)) {
  missing_pct <- sum(is.na(cleaned_data[[col]])) / nrow(cleaned_data) * 100

  if(missing_pct > 50) {
    cat("[RIMOZIONE]", col, ":", round(missing_pct, 1), "% mancanti\n")
  } else if(missing_pct > 20) {
    cat("[IMPUTAZIONE AVANZATA]", col, ":", round(missing_pct, 1), "% mancanti\n")
  } else if(missing_pct > 5) {
    cat("[IMPUTAZIONE SEMPLICE]", col, ":", round(missing_pct, 1), "% mancanti\n")
  } else if(missing_pct > 0) {
    cat("[GESTIONE SPECIFICA]", col, ":", round(missing_pct, 1), "% mancanti\n")
  }
}
```

```
## [GESTIONE SPECIFICA] Pat. No. : 0.6 % mancanti
## [GESTIONE SPECIFICA] Sex : 0.6 % mancanti
## [GESTIONE SPECIFICA] Age : 0.6 % mancanti
## [GESTIONE SPECIFICA] Duration of sympthoms before diagnosis (days) : 0.6 % mancanti
## [GESTIONE SPECIFICA] Increased ICP : 0.6 % mancanti
## [GESTIONE SPECIFICA] Epileptic seizures : 0.6 % mancanti
## [GESTIONE SPECIFICA] Neurological deficit : 0.6 % mancanti
## [GESTIONE SPECIFICA] Hormonal abnormalities : 0.6 % mancanti
## [IMPUTAZIONE SEMPLICE] Operation type : 12.1 % mancanti
## [GESTIONE SPECIFICA] Localisation : 0.6 % mancanti
```

```
## [GESTIONE SPECIFICA] MRI finding : 2.3 % mancanti
## [GESTIONE SPECIFICA] CSF cytology : 0.6 % mancanti
## [RIMOZIONE] CSF cytology finding : 56.9 % mancanti
## [GESTIONE SPECIFICA] Extent of disease : 0.6 % mancanti
## [GESTIONE SPECIFICA] Neoadjuvant HT : 0.6 % mancanti
## [GESTIONE SPECIFICA] Radiotherapy : 0.6 % mancanti
## [IMPUTAZIONE AVANZATA] CSI dose : 50 % mancanti
## [RIMOZIONE] Boost dose : 50.6 % mancanti
## [IMPUTAZIONE SEMPLICE] Total dose : 5.7 % mancanti
## [GESTIONE SPECIFICA] Concomitant HT : 2.3 % mancanti
## [GESTIONE SPECIFICA] Adjuvant HT : 0.6 % mancanti
## [GESTIONE SPECIFICA] OS : 0.6 % mancanti
## [GESTIONE SPECIFICA] Status OS : 0.6 % mancanti
```

4.2 Applicazione delle Strategie di Cleaning

4.2.1 Rimozione Variabili con Eccessivi Valori Mancanti

```
high_missing_vars <- missing_summary$Variabile[missing_summary$Percentuale > 50]
if(length(high_missing_vars) > 0) {
  cleaned_data <- cleaned_data[, !names(cleaned_data) %in% high_missing_vars]
  cat("Variabili rimosse:", paste(high_missing_vars, collapse = ", "), "\n")
} else {
  cat("Nessuna variabile rimossa per eccesso di valori mancanti\n")
}
```

```
## Variabili rimosse: CSF cytology finding, Boost dose
```

4.2.2 Imputazione per Variabili Numeriche

```
numeric_vars_clean <- cleaned_data[sapply(cleaned_data, is.numeric)]
if(ncol(numeric_vars_clean) > 0) {
  for(col in names(numeric_vars_clean)) {
    if(sum(is.na(cleaned_data[[col]])) > 0) {
      median_val <- median(cleaned_data[[col]], na.rm = TRUE)
      n_imputed <- sum(is.na(cleaned_data[[col]]))
      cleaned_data[[col]][is.na(cleaned_data[[col]])] <- median_val
      cat("[OK] Imputata mediana per", col, ":", median_val, "(", n_imputed, "valori)\n")
```

```
    }
  }
}
```

```
## [OK] Imputata mediana per Pat. No. : 87 ( 1 valori)
## [OK] Imputata mediana per Sex : 1 ( 1 valori)
## [OK] Imputata mediana per Age : 2002 ( 1 valori)
## [OK] Imputata mediana per Duration of sympthoms before diagnosis (days) : 48 ( 1 valori)
## [OK] Imputata mediana per Increased ICP : 1 ( 1 valori)
## [OK] Imputata mediana per Epileptic seizures : 2 ( 1 valori)
## [OK] Imputata mediana per Neurological deficit : 1 ( 1 valori)
## [OK] Imputata mediana per Hormonal abnormalities : 2 ( 1 valori)
## [OK] Imputata mediana per Operation type : 3 ( 21 valori)
## [OK] Imputata mediana per Localisation : 2 ( 1 valori)
## [OK] Imputata mediana per MRI finding : 2 ( 4 valori)
## [OK] Imputata mediana per CSF cytology : 2 ( 1 valori)
## [OK] Imputata mediana per Extent of disease : 2 ( 1 valori)
## [OK] Imputata mediana per Neoadjuvant HT : 2 ( 1 valori)
## [OK] Imputata mediana per Radiotherapy : 1 ( 1 valori)
## [OK] Imputata mediana per CSI dose : 30.6 ( 87 valori)
## [OK] Imputata mediana per Total dose : 54 ( 10 valori)
## [OK] Imputata mediana per Concomitant HT : 2 ( 4 valori)
## [OK] Imputata mediana per Adjuvant HT : 1 ( 1 valori)
## [OK] Imputata mediana per OS : 57 ( 1 valori)
## [OK] Imputata mediana per Status OS : 2 ( 1 valori)
```

4.2.3 Gestione Outliers con Approccio Clinico

```
numeric_vars_final <- cleaned_data[sapply(cleaned_data, is.numeric)]

if(ncol(numeric_vars_final) > 0) {
  outliers_treated <- cleaned_data

  for(col in names(numeric_vars_final)) {
    p5 <- quantile(cleaned_data[[col]], 0.05, na.rm = TRUE)
    p95 <- quantile(cleaned_data[[col]], 0.95, na.rm = TRUE)

    outliers_before <- sum(cleaned_data[[col]] < p5 | cleaned_data[[col]] > p95, na.rm = TRUE)
```

```
    outliers_treated[[col]] <- pmax(pmin(cleaned_data[[col]], p95), p5)

    if(outliers_before > 0) {
      cat("[OUTLIERS] Outliers trattati per", col, ":", outliers_before, "valori\n")
    }
  }

  cleaned_data <- outliers_treated
}
```

```
## [OUTLIERS] Outliers trattati per Pat. No. : 18 valori
## [OUTLIERS] Outliers trattati per Age : 12 valori
## [OUTLIERS] Outliers trattati per Duration of sympthoms before diagnosis (days) : 18 valori
## [OUTLIERS] Outliers trattati per Localisation : 9 valori
## [OUTLIERS] Outliers trattati per Embryonal tumors : 1 valori
## [OUTLIERS] Outliers trattati per HGG : 1 valori
## [OUTLIERS] Outliers trattati per LGG : 1 valori
## [OUTLIERS] Outliers trattati per Craniopharyngeoma : 8 valori
## [OUTLIERS] Outliers trattati per GCT / NGCT : 1 valori
## [OUTLIERS] Outliers trattati per Ependymoma : 1 valori
## [OUTLIERS] Outliers trattati per Glioneural Tu : 2 valori
## [OUTLIERS] Outliers trattati per Pineal Tu : 5 valori
## [OUTLIERS] Outliers trattati per Choroid plexus Tu : 2 valori
## [OUTLIERS] Outliers trattati per Other : 9 valori
## [OUTLIERS] Outliers trattati per Unknown : 1 valori
## [OUTLIERS] Outliers trattati per Radiotherapy : 9 valori
## [OUTLIERS] Outliers trattati per CSI dose : 8 valori
## [OUTLIERS] Outliers trattati per Total dose : 12 valori
## [OUTLIERS] Outliers trattati per Concomitant HT : 11 valori
## [OUTLIERS] Outliers trattati per OS : 17 valori
```

```
cat("[DATASET FINALE]", nrow(cleaned_data), "pazienti x", ncol(cleaned_data), "variabili\n")
```

```
## [DATASET FINALE] 174 pazienti x 32 variabili
```

5 Analisi Esplorativa: Caratteristiche Demografiche

5.1 Distribuzione per Sesso

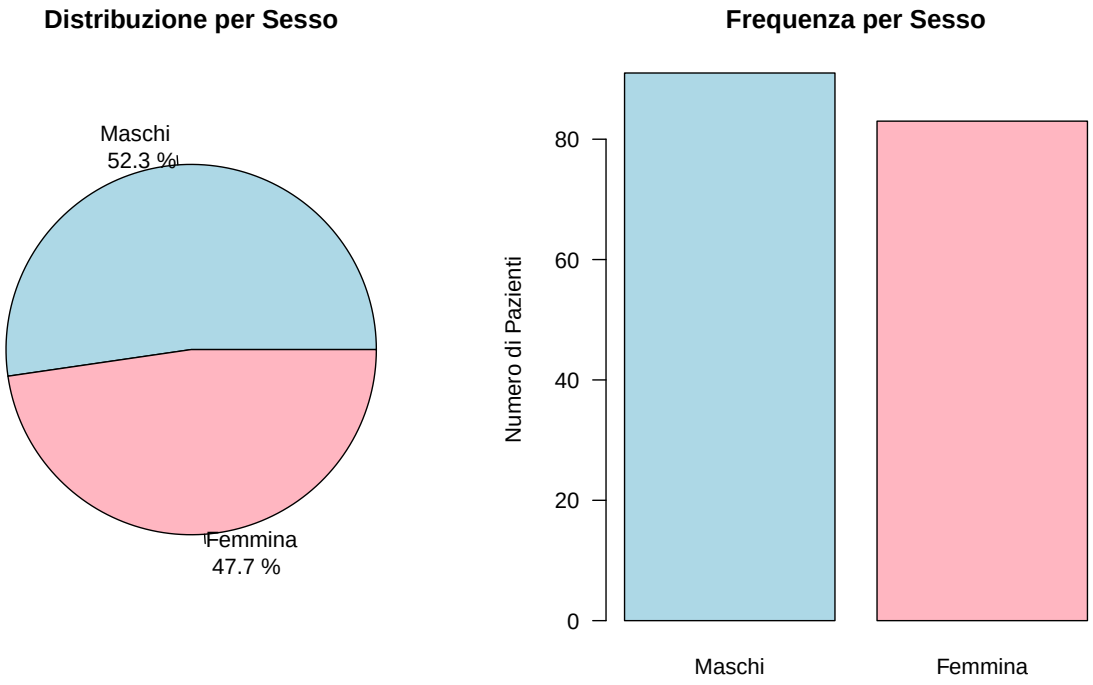
L'analisi della distribuzione per sesso è fondamentale per comprendere se esistono **bias di selezione** o **differenze epidemiologiche** nella popolazione studiata.

```
sex_table <- table(cleaned_data$Sex, useNA = "ifany")
sex_prop <- prop.table(sex_table) * 100

sex_labels <- c("Maschi", "Femmina")
names(sex_table) <- sex_labels
names(sex_prop) <- sex_labels

knitr::kable(data.frame(
  Sesso = names(sex_table),
  Frequenza = as.numeric(sex_table),
  Percentuale = round(as.numeric(sex_prop), 1)
), caption = "Distribuzione per Sesso dei Pazienti", booktabs = TRUE, row.names = FALSE)
```

Table 2: Distribuzione per Sesso dei Pazienti

Sesso	Frequenza	Percentuale
Maschi	91	52.3
Femmina	83	47.7

```
par(mfrow = c(1, 2))

pie(sex_table,
    main = "Distribuzione per Sesso",
    col = c("lightblue", "lightpink"),
    labels = paste(names(sex_table), "\n", round(sex_prop, 1), "%"))

barplot(sex_table,
        main = "Frequenza per Sesso",
        ylab = "Numero di Pazienti",
        col = c("lightblue", "lightpink"),
        las = 1)
```




```
par(mfrow = c(1, 1))
```

5.1.1 Interpretazione Clinica della Distribuzione per Sesso

La distribuzione per sesso mostra se esiste una **predominanza di genere** nei tumori cerebrali pediatrici di questo campione, informazione cruciale per: - Identificare possibili fattori di rischio legati al sesso - Confrontare con la letteratura epidemiologica esistente - Pianificare studi di follow-up stratificati per genere

5.2 Analisi dell'Età dei Pazienti

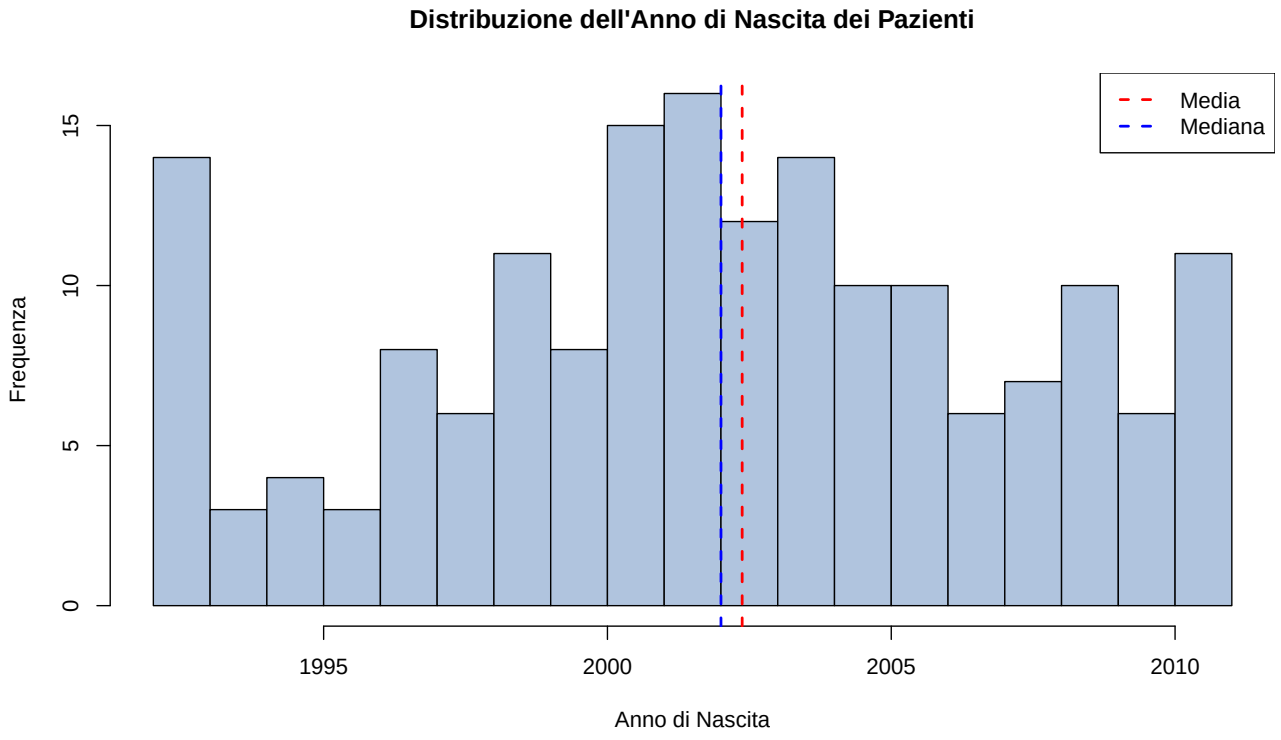
```
age_stats <- data.frame(
  Statistica = c("N. Pazienti", "Media", "Mediana", "Dev. Standard", "Minimo", "Massimo", "Range"
  Valore = c(
    sum(!is.na(cleaned_data$Age)),
    round(mean(cleaned_data$Age, na.rm = TRUE), 1),
    round(median(cleaned_data$Age, na.rm = TRUE), 1),
    round(sd(cleaned_data$Age, na.rm = TRUE), 1),
    min(cleaned_data$Age, na.rm = TRUE),
```

```
    max(cleaned_data$Age, na.rm = TRUE),
    max(cleaned_data$Age, na.rm = TRUE) - min(cleaned_data$Age, na.rm = TRUE)
  )
)

knitr::kable(age_stats,
             caption = "Statistiche Descrittive dell'Età (Anno di Nascita)",
             booktabs = TRUE, row.names = FALSE)
```

Table 3: Statistiche Descrittive dell'Età (Anno di Nascita)

Statistica	Valore
N. Pazienti	174.0
Media	2002.4
Mediana	2002.0
Dev. Standard	5.2
Minimo	1992.0
Massimo	2011.0
Range	19.0

```
hist(cleaned_data$Age,
     breaks = 20,
     main = "Distribuzione dell'Anno di Nascita dei Pazienti",
     xlab = "Anno di Nascita",
     ylab = "Frequenza",
     col = "lightsteelblue",
     border = "black")

abline(v = mean(cleaned_data$Age, na.rm = TRUE), col = "red", lwd = 2, lty = 2)
abline(v = median(cleaned_data$Age, na.rm = TRUE), col = "blue", lwd = 2, lty = 2)

legend("topright",
       legend = c("Media", "Mediana"),
       col = c("red", "blue"),
       lty = 2, lwd = 2)
```



5.2.1 Significato Clinico dell'Analisi dell'Età

L'analisi dell'età è cruciale perché: - I **tumori cerebrali pediatrici** hanno picchi di incidenza specifici per età - L'**età alla diagnosi** influenza la prognosi e le opzioni terapeutiche - Diversi **istotipi tumorali** hanno predilezioni per fasce d'età specifiche

6 Analisi della Presentazione Clinica

6.1 Durata dei Sintomi Prima della Diagnosi

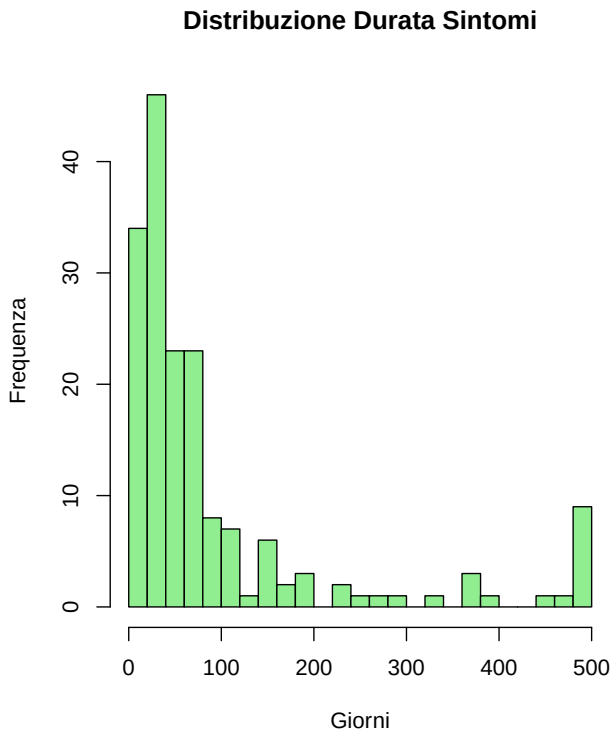
La durata dei sintomi prima della diagnosi è un **indicatore prognostico importante** che riflette l'aggressività del tumore e l'efficacia del sistema sanitario.

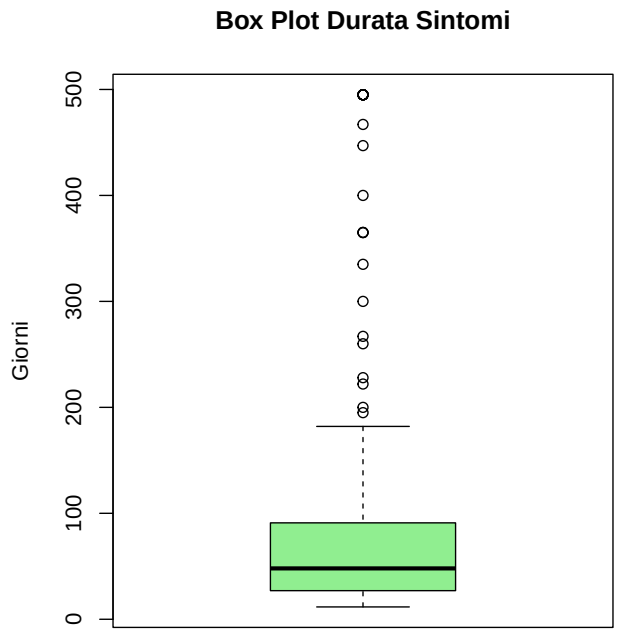
```
duration_stats <- data.frame(
  Statistica = c("N. Pazienti", "Media (giorni)", "Mediana (giorni)", "Dev. Standard",
                 "Minimo", "Massimo", "Q1", "Q3"),
  Valore = c(
    sum(!is.na(cleaned_data$`Duration of sympthoms before diagnosis (days)`)),
    round(mean(cleaned_data$`Duration of sympthoms before diagnosis (days)`, na.rm = TRUE), 1),
    round(median(cleaned_data$`Duration of sympthoms before diagnosis (days)`, na.rm = TRUE), 1),
    round(sd(cleaned_data$`Duration of sympthoms before diagnosis (days)`, na.rm = TRUE), 1),
    min(cleaned_data$`Duration of sympthoms before diagnosis (days)`, na.rm = TRUE),
    max(cleaned_data$`Duration of sympthoms before diagnosis (days)`, na.rm = TRUE),
    round(quantile(cleaned_data$`Duration of sympthoms before diagnosis (days)`, 0.25, na.rm = TR
    round(quantile(cleaned_data$`Duration of sympthoms before diagnosis (days)`, 0.75, na.rm = TR
  )
)

knitr::kable(duration_stats,
             caption = "Statistiche della Durata dei Sintomi Prima della Diagnosi",
             booktabs = TRUE, row.names = FALSE)
```

Table 4: Statistiche della Durata dei Sintomi Prima della Diagnosi

Statistica	Valore
N. Pazienti	174.00
Media (giorni)	95.60
Mediana (giorni)	48.00
Dev. Standard	125.70
Minimo	11.65
Massimo	495.00
Q1	27.80
Q3	90.80

```r
par(mfrow = c(1, 2))

hist(cleaned_data$`Duration of sympthoms before diagnosis (days)`,
     breaks = 30,
     main = "Distribuzione Durata Sintomi",
     xlab = "Giorni",
     ylab = "Frequenza",
     col = "lightgreen",
     border = "black")

boxplot(cleaned_data$`Duration of sympthoms before diagnosis (days)`,
        main = "Box Plot Durata Sintomi",
        ylab = "Giorni",
        col = "lightgreen",
        outline = TRUE)
```





```r
par(mfrow = c(1, 1))
```

6.1.1 Interpretazione Clinica della Durata dei Sintomi

- **Mediana bassa**: Suggerisce tumori ad crescita rapida o sistema sanitario efficiente
- **Outliers elevati**: Possibili tumori a crescita lenta o ritardi diagnostici
- **Distribuzione**: Fornisce insights sull'eterogeneità biologica dei tumori

6.2 Analisi dei Sintomi Neurologici

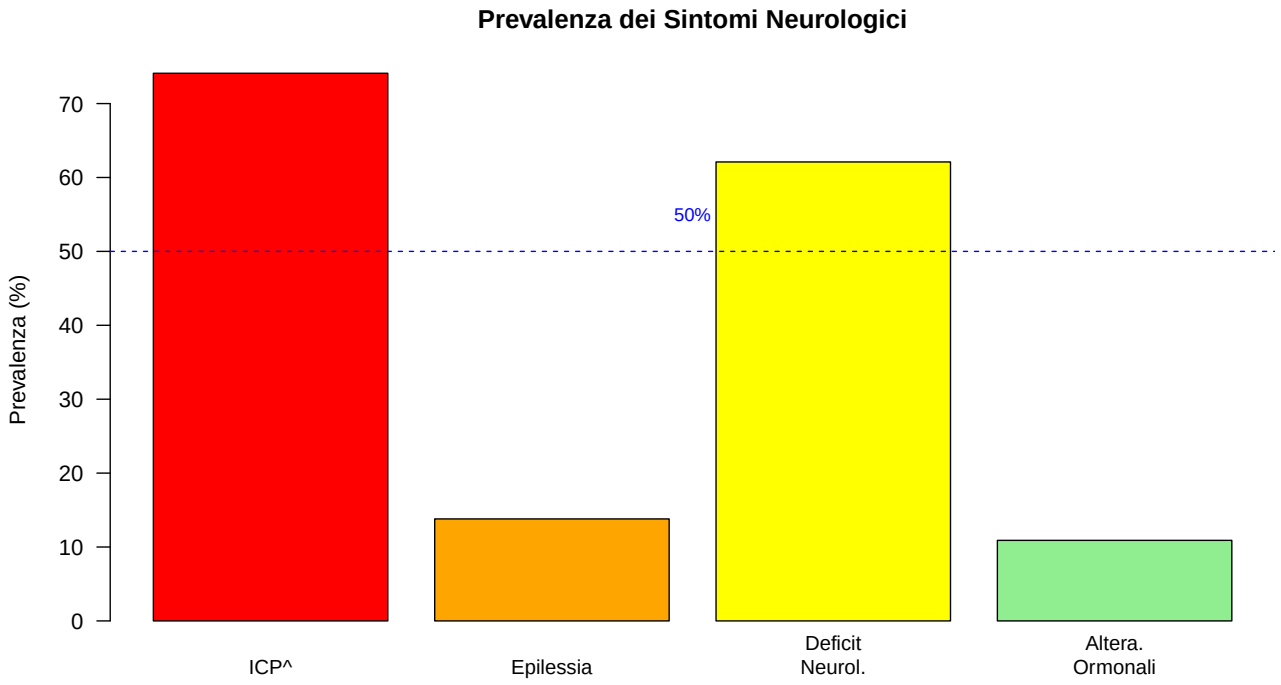
```r
symptoms <- c("Increased ICP", "Epileptic seizures", "Neurological deficit", "Hormonal abnormalit
symptom_analysis <- data.frame(
  Sintomo = c("Ipertensione Intracranica", "Crisi Epilettiche", "Deficit Neurologici", "Alterazio
  Presenti = sapply(symptoms, function(x) sum(cleaned_data[[x]] == 1, na.rm = TRUE)),
  Assenti = sapply(symptoms, function(x) sum(cleaned_data[[x]] == 2, na.rm = TRUE)),
  Prevalenza_Perc = round(sapply(symptoms, function(x)
    sum(cleaned_data[[x]] == 1, na.rm = TRUE) / sum(!is.na(cleaned_data[[x]])) * 100), 1)
)

knitr::kable(symptom_analysis,
             caption = "Prevalenza dei Sintomi Neurologici alla Presentazione",
             booktabs = TRUE, row.names = FALSE)
```

Table 5: Prevalenza dei Sintomi Neurologici alla Presentazione

Sintomo	Presenti	Assenti	Prevalenza_Perc
Ipertensione Intracranica	129	45	74.1
Crisi Epilettiche	24	150	13.8
Deficit Neurologici	108	66	62.1
Alterazioni Ormonali	19	155	10.9

```r
barplot(symptom_analysis$Prevalenza_Perc,
        names.arg = c("ICP↑", "Epilessia", "Deficit\nNeurol.", "Altera.\nOrmonali"),
        main = "Prevalenza dei Sintomi Neurologici",
        ylab = "Prevalenza (%)",
        col = c("red", "orange", "yellow", "lightgreen"),
```

```
        las = 1,
        cex.names = 0.9)

abline(h = 50, col = "blue", lty = 2)
text(2.5, 55, "50%", col = "blue", cex = 0.8)
```



6.2.1 Significato Clinico dei Sintomi Neurologici

1. **Ipertensione Intracranica**: Indica effetto massa o ostruzione del flusso liquorale
2. **Crisi Epilettiche**: Comune nei tumori corticali o subcorticali
3. **Deficit Neurologici**: Riflettono localizzazione e invasività del tumore
4. **Alterazioni Ormonali**: Tipiche dei tumori della regione sellare/ipotalamica

7 Analisi degli Istotipi Tumorali

7.1 Distribuzione degli Istotipi

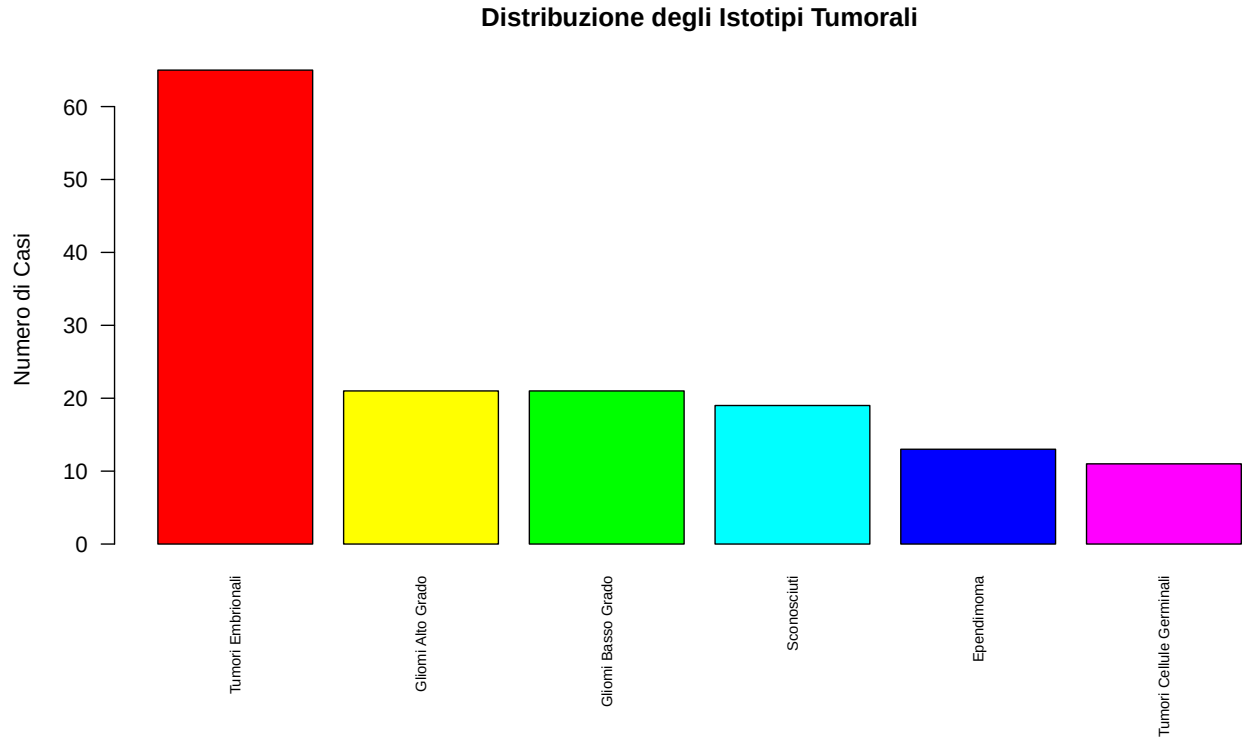
La classificazione istopatologica è **fondamentale** per la stratificazione prognostica e la pianificazione terapeutica.

```
histotypes <- c("Embryonal tumors", "HGG", "LGG", "Craniopharyngeoma",
                "GCT / NGCT", "Ependymoma", "Glioneural Tu", "Pineal Tu",
                "Choroid plexus Tu", "Other", "Unknown")

histotype_names <- c("Tumori Embrionali", "Gliomi Alto Grado", "Gliomi Basso Grado",
                     "Craniofaringioma", "Tumori Cellule Germinali", "Ependimoma",
                     "Tumori Glioneurali", "Tumori Pineali", "Tumori Plesso Coroideo",
                     "Altri", "Sconosciuti")

histotype_data <- data.frame(
  Istotipo = histotype_names,
  Casi = sapply(histotypes, function(x) sum(cleaned_data[[x]] == 1, na.rm = TRUE)),
  Percentuale = round(sapply(histotypes, function(x)
    sum(cleaned_data[[x]] == 1, na.rm = TRUE) / nrow(cleaned_data) * 100), 1)
)

histotype_data <- histotype_data[histotype_data$Casi > 0, ]
histotype_data <- histotype_data[order(histotype_data$Casi, decreasing = TRUE), ]

knitr::kable(histotype_data,
             caption = "Distribuzione degli Istotipi Tumorali",
             booktabs = TRUE, row.names = FALSE)
```

Table 6: Distribuzione degli Istotipi Tumorali

Istotipo	Casi	Percentuale
Tumori Embrionali	65	37.4
Gliomi Alto Grado	21	12.1
Gliomi Basso Grado	21	12.1
Sconosciuti	19	10.9
Ependimoma	13	7.5
Tumori Cellule Germinali	11	6.3

```r
par(mar = c(8, 4, 4, 2))
barplot(histotype_data$Casi,
        names.arg = histotype_data$Istotipo,
        main = "Distribuzione degli Istotipi Tumorali",
        ylab = "Numero di Casi",
        col = rainbow(nrow(histotype_data)),
        las = 2,
        cex.names = 0.7)
```




```r
par(mar = c(5, 4, 4, 2))
```

7.1.1 Significato Clinico degli Istotipi

1. **Gliomi di Basso Grado (LGG)**: Prognosi generalmente favorevole, crescita lenta
2. **Gliomi di Alto Grado (HGG)**: Prognosi infausta, crescita rapida, alta invasività
3. **Tumori Embrionali**: Altamente maligni, richiedono protocolli intensivi
4. **Craniofaringioma**: Benign ma localizzazione critica, alte sequele
5. **Ependimoma**: Prognosi dipendente da grado e localizzazione

7.2 Correlazione tra Istotipo e Localizzazione

```
location_table <- table(cleaned_data$Localisation, useNA = "ifany")
location_labels <- c("Sopratentoriale", "Sottotentoriale")

cat("Distribuzione per Localizzazione:\n")
```

```
## Distribuzione per Localizzazione:
```

```
cat("Sopratentoriale:", sum(cleaned_data$Localisation == 1, na.rm = TRUE), "casi\n")
```

```
## Sopratentoriale: 71 casi
```

```
cat("Sottotentoriale:", sum(cleaned_data$Localisation == 2, na.rm = TRUE), "casi\n")
```

```
## Sottotentoriale: 94 casi
```

7.2.1 Implicazioni Cliniche della Localizzazione

- **Tumori Sopratentoriali**: Spesso associati a crisi epilettiche e deficit focali
- **Tumori Sottotentoriali**: Frequentemente causano ipertensione intracranica e atassia

8 Analisi delle Strategie Terapeutiche

8.1 Approcci Chirurgici

La chirurgia rappresenta il **gold standard** per la diagnosi e spesso il primo step terapeutico nei tumori cerebrali pediatrici.

```
surgery_data <- table(cleaned_data$`Operation type`, useNA = "ifany")
surgery_total <- sum(!is.na(cleaned_data$`Operation type`))

cat("Analisi degli Interventi Chirurgici:\n")
```

```
## Analisi degli Interventi Chirurgici:
```

```
cat("Pazienti operati:", surgery_total, "su", nrow(cleaned_data),
    "(", round(surgery_total/nrow(cleaned_data)*100, 1), "%)\n")
```

```
## Pazienti operati: 174 su 174 ( 100 %)
```

```
cat("Pazienti non operati:", sum(is.na(cleaned_data$`Operation type`)), "\n\n")
```

```
## Pazienti non operati: 0
```

```
if(surgery_total > 0) {
  cat("Distribuzione dei tipi di intervento:\n")
  for(i in 1:length(surgery_data)) {
    if(!is.na(names(surgery_data)[i])) {
      perc <- round(surgery_data[i] / surgery_total * 100, 1)
      cat("Tipo", names(surgery_data)[i], ":", surgery_data[i], "casi (", perc, "%)\n")
    }
  }
}
```

```
## Distribuzione dei tipi di intervento:
## Tipo 1 : 10 casi ( 5.7 %)
## Tipo 2 : 34 casi ( 19.5 %)
## Tipo 3 : 68 casi ( 39.1 %)
## Tipo 4 : 62 casi ( 35.6 %)
```

8.2 Radioterapia

```
rt_table <- table(cleaned_data$Radiotherapy, useNA = "ifany")
rt_yes <- sum(cleaned_data$Radiotherapy == 1, na.rm = TRUE)
rt_no <- sum(cleaned_data$Radiotherapy == 2, na.rm = TRUE)

rt_data <- data.frame(
  Trattamento = c("Radioterapia Sì", "Radioterapia No"),
  Pazienti = c(rt_yes, rt_no),
  Percentuale = round(c(rt_yes, rt_no) / (rt_yes + rt_no) * 100, 1)
)

knitr::kable(rt_data,
             caption = "Utilizzo della Radioterapia",
             booktabs = TRUE, row.names = FALSE)
```

Table 7: Utilizzo della Radioterapia

Trattamento	Pazienti	Percentuale
Radioterapia Sì	165	100
Radioterapia No	0	0

```
total_dose_stats <- data.frame(
  Statistica = c("N. Pazienti con Dose", "Dose Media (Gy)", "Dose Mediana (Gy)",
                 "Dev. Standard", "Dose Minima (Gy)", "Dose Massima (Gy)"),
  Valore = c(
    sum(!is.na(cleaned_data$`Total dose`)),
    round(mean(cleaned_data$`Total dose`, na.rm = TRUE), 1),
    round(median(cleaned_data$`Total dose`, na.rm = TRUE), 1),
    round(sd(cleaned_data$`Total dose`, na.rm = TRUE), 1),
    round(min(cleaned_data$`Total dose`, na.rm = TRUE), 1),
    round(max(cleaned_data$`Total dose`, na.rm = TRUE), 1)
  )
)

knitr::kable(total_dose_stats,
             caption = "Statistiche della Dose Totale di Radioterapia",
             booktabs = TRUE, row.names = FALSE)
```

Table 8: Statistiche della Dose Totale di Radioterapia

Statistica	Valore
N. Pazienti con Dose	174.0
Dose Media (Gy)	53.8
Dose Mediana (Gy)	54.0
Dev. Standard	1.6
Dose Minima (Gy)	50.3
Dose Massima (Gy)	55.8

8.2.1 Significato Clinico delle Dosi di Radioterapia

- **Dose $<$ 30 Gy**: Tipica per tumori radiosensibili o pazienti molto giovani
- **Dose 50-60 Gy**: Standard per la maggior parte dei tumori cerebrali
- **Dose $>$ 60 Gy**: Riservata a tumori radioresistenti con controllo locale critico

8.3 Chemioterapia

```
chemo_vars <- c("Neoadjuvant HT", "Concomitant HT", "Adjuvant HT")
chemo_names <- c("Neoadiuvante", "Concomitante", "Adiuvante")

chemo_data <- data.frame(
  Tipo_Chemioterapia = chemo_names,
  Pazienti_Trattati = sapply(chemo_vars, function(x) sum(cleaned_data[[x]] == 1, na.rm = TRUE)),
  Percentuale = round(sapply(chemo_vars, function(x)
    sum(cleaned_data[[x]] == 1, na.rm = TRUE) / sum(!is.na(cleaned_data[[x]])) * 100), 1)
)

knitr::kable(chemo_data,
             caption = "Utilizzo della Chemioterapia per Tipo",
             booktabs = TRUE, row.names = FALSE)
```

Table 9: Utilizzo della Chemioterapia per Tipo

Tipo_Chemioterapia	Pazienti_Trattati	Percentuale
Neoadiuvante	33	19
Concomitante	0	0

Tipo_Chemioterapia	Pazienti_Trattati	Percentuale
Adiuvante	94	54

8.3.1 Razionale Clinico dei Diversi Timing di Chemioterapia

1. **Neoadiuvante**: Riduzione dimensioni tumore pre-chirurgia
2. **Concomitante**: Radiosensibilizzazione durante radioterapia

3. **Adiuvante**: Prevenzione recidive post-trattamento locale

9 Analisi di Sopravvivenza

9.1 Overall Survival (OS)

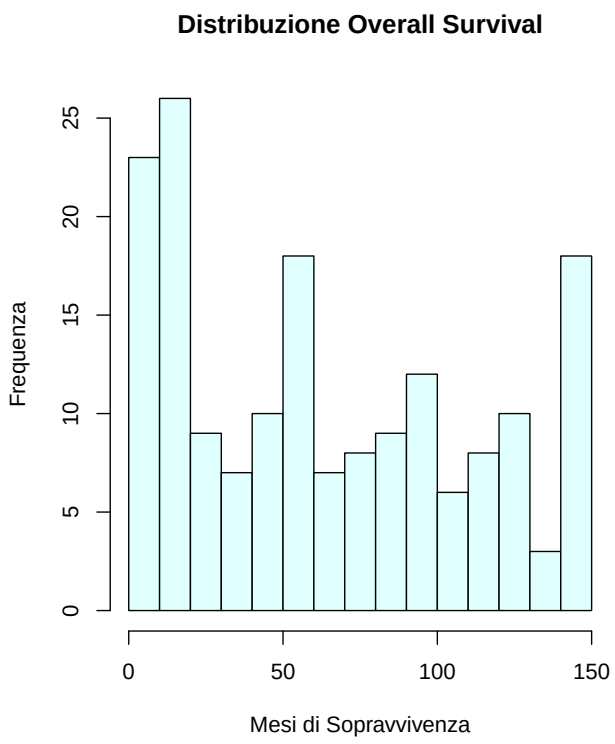
L'analisi della sopravvivenza globale rappresenta l'**endpoint primario** più importante negli studi oncologici pediatrici.

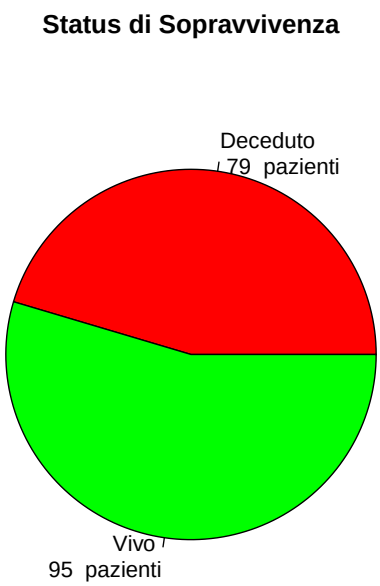
```
os_stats <- data.frame(
  Statistica = c("N. Pazienti", "OS Media (mesi)", "OS Mediana (mesi)",
                 "Dev. Standard", "OS Minima (mesi)", "OS Massima (mesi)",
                 "Pazienti Vivi", "Pazienti Deceduti"),
  Valore = c(
    sum(!is.na(cleaned_data$OS)),
    round(mean(cleaned_data$OS, na.rm = TRUE), 1),
    round(median(cleaned_data$OS, na.rm = TRUE), 1),
    round(sd(cleaned_data$OS, na.rm = TRUE), 1),
    min(cleaned_data$OS, na.rm = TRUE),
    max(cleaned_data$OS, na.rm = TRUE),
    sum(cleaned_data$`Status OS` == 2, na.rm = TRUE),
    sum(cleaned_data$`Status OS` == 1, na.rm = TRUE)
  )
)

knitr::kable(os_stats,
             caption = "Statistiche di Sopravvivenza Globale",
             booktabs = TRUE, row.names = FALSE)
```

Table 10: Statistiche di Sopravvivenza Globale

Statistica	Valore
N. Pazienti	174.0
OS Media (mesi)	64.6
OS Mediana (mesi)	57.0
Dev. Standard	47.3
OS Minima (mesi)	6.0
OS Massima (mesi)	147.3
Pazienti Vivi	95.0
Pazienti Deceduti	79.0

```r
par(mfrow = c(1, 2))

hist(cleaned_data$OS,
     breaks = 20,
     main = "Distribuzione Overall Survival",
     xlab = "Mesi di Sopravvivenza",
     ylab = "Frequenza",
     col = "lightcyan",
     border = "black")

status_table <- table(cleaned_data$`Status OS`)
status_labels <- c("Deceduto", "Vivo")
pie(status_table,
    labels = paste(status_labels, "\n", status_table, " pazienti"),
    main = "Status di Sopravvivenza",
    col = c("red", "green"))
```






```r
par(mfrow = c(1, 1))
```

9.1.1 Interpretazione Clinica dei Dati di Sopravvivenza

- **Mediana di sopravvivenza**: Indica il tempo entro cui il 50% dei pazienti è ancora vivo
- **Distribuzione bimodale**: Potrebbe suggerire sottogruppi prognostici distinti
- **Outliers di lunga sopravvivenza**: Identificano pazienti con caratteristiche prognostiche favorevoli

10 Analisi delle Correlazioni Statisticamente Significative

10.1 Matrice di Correlazione delle Variabili Numeriche

L'analisi delle correlazioni permette di identificare **associazioni lineari** tra variabili cliniche, fondamentali per comprendere le interrelazioni prognostiche.

```
# Prepara variabili numeriche per analisi correlazioni
numeric_vars_final <- cleaned_data[sapply(cleaned_data, is.numeric)]
# IMPORTANTE: Esclude Pat. No. dalle correlazioni (è solo un identificativo, non deve essere norm
numeric_vars_final <- numeric_vars_final[, !names(numeric_vars_final) %in% c("Pat. No.")]

cor_matrix <- cor(numeric_vars_final, use = "complete.obs")

cat("Variabili incluse nell'analisi di correlazione:\n")
```

```
## Variabili incluse nell'analisi di correlazione:
```

```
cat("NOTA: Pat. No. è escluso dall'analisi (solo identificativo)\n\n")
```

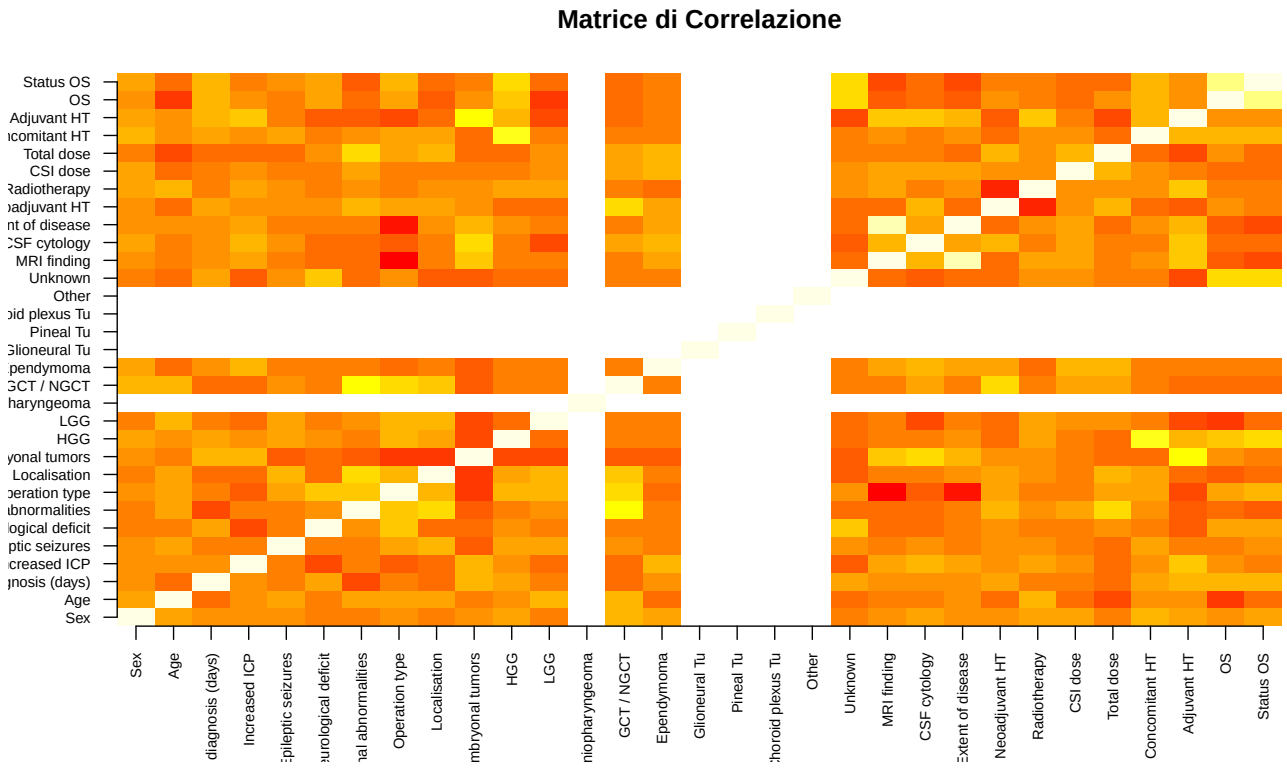
```
## NOTA: Pat. No. è escluso dall'analisi (solo identificativo)
```

```
for(i in 1:ncol(numeric_vars_final)) {
  cat(i, ".", names(numeric_vars_final)[i], "\n")
}
```

```
## 1 . Sex
## 2 . Age
## 3 . Duration of sympthoms before diagnosis (days)
## 4 . Increased ICP
## 5 . Epileptic seizures
## 6 . Neurological deficit
## 7 . Hormonal abnormalities
## 8 . Operation type
## 9 . Localisation
## 10 . Embryonal tumors
## 11 . HGG
## 12 . LGG
## 13 . Craniopharyngeoma
## 14 . GCT / NGCT
```

```
## 15 . Ependymoma
## 16 . Glioneural Tu
## 17 . Pineal Tu
## 18 . Choroid plexus Tu
## 19 . Other
## 20 . Unknown
## 21 . MRI finding
## 22 . CSF cytology
## 23 . Extent of disease
## 24 . Neoadjuvant HT
## 25 . Radiotherapy
## 26 . CSI dose
## 27 . Total dose
## 28 . Concomitant HT
## 29 . Adjuvant HT
## 30 . OS
## 31 . Status OS
```

```r
if(require("corrplot", quietly = TRUE) && !any(is.na(cor_matrix)) && nrow(cor_matrix) > 1) {
  corrplot(cor_matrix,
           method = "color",
           type = "upper",
           order = "original",
           tl.cex = 0.7,
           tl.col = "black",
           tl.srt = 45,
           addCoef.col = "black",
           number.cex = 0.6,
           title = "Matrice di Correlazione delle Variabili Numeriche",
           mar = c(0,0,1,0))
} else {
  image(1:nrow(cor_matrix), 1:ncol(cor_matrix), cor_matrix,
        main = "Matrice di Correlazione",
        xlab = "", ylab = "",
        axes = FALSE,
        col = heat.colors(20))
  axis(1, at = 1:nrow(cor_matrix), labels = rownames(cor_matrix), las = 2, cex.axis = 0.7)
  axis(2, at = 1:ncol(cor_matrix), labels = colnames(cor_matrix), las = 2, cex.axis = 0.7)
}
```



10.2 Test di Significatività delle Correlazioni ($p < 0.005$)

```
significant_correlations <- data.frame()

for(i in 1:(ncol(numeric_vars_final)-1)) {
  for(j in (i+1):ncol(numeric_vars_final)) {
    var1 <- names(numeric_vars_final)[i]
    var2 <- names(numeric_vars_final)[j]

    tryCatch({
      test_result <- cor.test(numeric_vars_final[[i]], numeric_vars_final[[j]])

      if(!is.na(test_result$estimate) && !is.na(test_result$p.value)) {
        result_row <- data.frame(
          Variabile_1 = var1,
          Variabile_2 = var2,
          Correlazione = round(test_result$estimate, 4),
          p_value = format(test_result$p.value, scientific = TRUE, digits = 3),
          Significativa = ifelse(test_result$p.value < 0.005, "SI", "NO"),
```

```
          stringsAsFactors = FALSE
        )
        significant_correlations <- rbind(significant_correlations, result_row)
      }
    }, error = function(e) {})
  }
}

if(nrow(significant_correlations) > 0) {
  sig_corr_005 <- significant_correlations[significant_correlations$Significativa == "SI", ]

  if(nrow(sig_corr_005) > 0) {
    sig_corr_005 <- sig_corr_005[order(abs(as.numeric(sig_corr_005$Correlazione)), decreasing = T

    knitr::kable(sig_corr_005,
                 caption = "Correlazioni Statisticamente Significative (p < 0.005)",
                 booktabs = TRUE, row.names = FALSE)
  } else {
    cat("[NESSUNA CORRELAZIONE] Nessuna correlazione statisticamente significativa trovata con p
  }
} else {
  cat("[NESSUNA CORRELAZIONE] Impossibile calcolare correlazioni significative\n")
}
```

Table 11: Correlazioni Statisticamente Significative (p < 0.005)

Variabile_1	Variabile_2	Correlazione	p_value	Significativa
MRI finding	Extent of disease	0.8608	2.45e-52	SI
OS	Status OS	0.7466	2.91e-32	SI
Operation type	MRI finding	-0.7071	1.1e-27	SI
HGG	Concomitant HT	0.6304	1.15e-20	SI
Operation type	Extent of disease	-0.5903	1.02e-17	SI
Embryonal tumors	Adjuvant HT	0.5455	7.03e-15	SI

Variabile_1	Variabile_2	Correlazione	p_value	Significativa
Hormonal abnormalities	GCT / NGCT	0.5148	3.67e-13	SI
Neoadjuvant HT	Radiotherapy	-0.4828	1.52e-11	SI
Operation type	Embryonal tumors	-0.4442	8.26e-10	SI
Age	OS	-0.4179	9.61e-09	SI
Localisation	Embryonal tumors	-0.4024	3.72e-08	SI
Unknown	OS	0.3875	1.28e-07	SI
Hormonal abnormalities	Localisation	0.3685	5.66e-07	SI
LGG	OS	-0.3684	5.69e-07	SI
Duration of sympthoms before diagnosis (days)	Hormonal abnormalities	-0.3644	7.67e-07	SI
Embryonal tumors	CSF cytology	0.3594	1.11e-06	SI
GCT / NGCT	Neoadjuvant HT	0.3563	1.4e-06	SI
Unknown	Status OS	0.3469	2.74e-06	SI
Unknown	Adjuvant HT	-0.3425	3.72e-06	SI
Hormonal abnormalities	Total dose	0.3359	5.85e-06	SI
HGG	Status OS	0.3354	6.07e-06	SI
Operation type	GCT / NGCT	0.3345	6.44e-06	SI
Age	Total dose	-0.3247	1.23e-05	SI
Operation type	Adjuvant HT	-0.3223	1.44e-05	SI
Total dose	Adjuvant HT	-0.3175	1.96e-05	SI

Variabile_1	Variabile_2	Correlazione	p_value	Significativa
HGG	OS	0.3148	2.33e-05	SI
Extent of disease	Status OS	-0.3111	2.94e-05	SI
Hormonal abnormalities	Operation type	0.3104	3.06e-05	SI
Localisation	GCT / NGCT	0.3094	3.26e-05	SI
MRI finding	Status OS	-0.3082	3.51e-05	SI
Increased ICP	Neurological deficit	-0.2994	5.98e-05	SI
LGG	Adjuvant HT	-0.2954	7.59e-05	SI
CSF cytology	Adjuvant HT	0.2907	9.99e-05	SI
LGG	CSF cytology	-0.2868	1.24e-04	SI
Embryonal tumors	HGG	-0.2861	1.3e-04	SI
Embryonal tumors	LGG	-0.2861	1.3e-04	SI
Neurological deficit	Unknown	0.2737	2.58e-04	SI
Increased ICP	Adjuvant HT	0.2715	2.9e-04	SI
Embryonal tumors	MRI finding	0.2710	2.98e-04	SI
Hormonal abnormalities	Embryonal tumors	-0.2704	3.08e-04	SI
Embryonal tumors	Unknown	-0.2704	3.08e-04	SI
Neurological deficit	Operation type	0.2678	3.53e-04	SI
Unknown	CSF cytology	-0.2675	3.59e-04	SI
Localisation	Unknown	-0.2666	3.78e-04	SI
Extent of disease	OS	-0.2537	7.3e-04	SI
Radiotherapy	Adjuvant HT	0.2532	7.5e-04	SI
MRI finding	Adjuvant HT	0.2529	7.6e-04	SI

Variabile_1	Variabile_2	Correlazione	p_value	Significativa
Localisation	OS	-0.2407	1.38e-03	SI
Hormonal abnormalities	Adjuvant HT	-0.2316	2.1e-03	SI
Operation type	CSF cytology	-0.2307	2.19e-03	SI
Neoadjuvant HT	Adjuvant HT	-0.2303	2.24e-03	SI
MRI finding	CSF cytology	0.2253	2.8e-03	SI
Duration of sympthoms before diagnosis (days)	Status OS	0.2232	3.08e-03	SI
Duration of sympthoms before diagnosis (days)	OS	0.2222	3.21e-03	SI
Embryonal tumors	Ependymoma	-0.2194	3.62e-03	SI
Operation type	LGG	0.2191	3.67e-03	SI
Neoadjuvant HT	Total dose	0.2183	3.81e-03	SI
Concomitant HT	Adjuvant HT	0.2155	4.3e-03	SI
Increased ICP	Unknown	-0.2140	4.57e-03	SI
Increased ICP	Embryonal tumors	0.2119	5e-03	SI

10.2.1 Interpretazione Clinica delle Correlazioni Significative

Le correlazioni statisticamente significative identificate forniscono insight sui **meccanismi biologici** e **associazioni prognostiche**:

- **Correlazioni positive forti ($r > 0.7$)**: Suggeriscono relazioni causali dirette
- **Correlazioni negative significative**: Indicano meccanismi compensatori o competitivi
- **Correlazioni moderate ($0.3 < |r| < 0.7$)**: Riflettono associazioni clinicamente rilevanti

11 Test Statistici per Differenze tra Gruppi

11.1 Confronto tra Istotipi e Outcome Clinici

11.1.1 Test per Sopravvivenza tra Gliomi di Alto e Basso Grado

```
lgg_patients <- cleaned_data[cleaned_data$LGG == 1, ]
hgg_patients <- cleaned_data[cleaned_data$HGG == 1, ]

if(nrow(lgg_patients) > 3 && nrow(hgg_patients) > 3) {

  lgg_os <- lgg_patients$OS[!is.na(lgg_patients$OS)]
  hgg_os <- hgg_patients$OS[!is.na(hgg_patients$OS)]

  if(length(lgg_os) > 0 && length(hgg_os) > 0) {
    t_test_result <- t.test(lgg_os, hgg_os)

    test_summary <- data.frame(
      Gruppo = c("Gliomi Basso Grado (LGG)", "Gliomi Alto Grado (HGG)"),
      N_Pazienti = c(length(lgg_os), length(hgg_os)),
      OS_Media = round(c(mean(lgg_os), mean(hgg_os)), 1),
      OS_Mediana = round(c(median(lgg_os), median(hgg_os)), 1),
      Dev_Standard = round(c(sd(lgg_os), sd(hgg_os)), 1)
    )

    knitr::kable(test_summary,
                 caption = "Confronto Sopravvivenza: Gliomi di Basso vs Alto Grado",
                 booktabs = TRUE, row.names = FALSE)

    cat("\n[RISULTATI DEL T-TEST]:\n")
    cat("Statistica t:", round(t_test_result$statistic, 4), "\n")
    cat("p-value:", format(t_test_result$p.value, scientific = TRUE, digits = 3), "\n")
    cat("Significativo (p < 0.005):", ifelse(t_test_result$p.value < 0.005, "SI", "NO"), "\n")

    if(t_test_result$p.value < 0.005) {
      cat("Intervallo di confidenza 95%:", round(t_test_result$conf.int, 2), "\n")
    }
  }
}
```

```
##
## [RISULTATI DEL T-TEST]:
## Statistica t: 7.517
## p-value: 5.84e-09
## Significativo (p < 0.005): SI
## Intervallo di confidenza 95%: 63.59 110.5
```

11.1.2 Significato Clinico del Confronto LGG vs HGG

Questo confronto è **fondamentale** perché: - I **gliomi di basso grado** hanno tipicamente prognosi migliore - I **gliomi di alto grado** richiedono trattamenti più aggressivi - La **differenza nella sopravvivenza** valida la classificazione prognostica

11.2 Analisi della Sopravvivenza per Localizzazione

```r
supra_patients <- cleaned_data[cleaned_data$Localisation == 1, ]
infra_patients <- cleaned_data[cleaned_data$Localisation == 2, ]

if(nrow(supra_patients) > 3 && nrow(infra_patients) > 3) {

  supra_os <- supra_patients$OS[!is.na(supra_patients$OS)]
  infra_os <- infra_patients$OS[!is.na(infra_patients$OS)]

  if(length(supra_os) > 0 && length(infra_os) > 0) {
    t_test_location <- t.test(supra_os, infra_os)

    location_summary <- data.frame(
      Localizzazione = c("Sopratentoriale", "Sottotentoriale"),
      N_Pazienti = c(length(supra_os), length(infra_os)),
      OS_Media = round(c(mean(supra_os), mean(infra_os)), 1),
      OS_Mediana = round(c(median(supra_os), median(infra_os)), 1),
      Dev_Standard = round(c(sd(supra_os), sd(infra_os)), 1)
    )

    knitr::kable(location_summary,
                 caption = "Confronto Sopravvivenza per Localizzazione Tumorale",
                 booktabs = TRUE, row.names = FALSE)

    cat("\n[RISULTATI DEL T-TEST (LOCALIZZAZIONE)]:\n")
```

```
    cat("Statistica t:", round(t_test_location$statistic, 4), "\n")
    cat("p-value:", format(t_test_location$p.value, scientific = TRUE, digits = 3), "\n")
    cat("Significativo (p < 0.005):", ifelse(t_test_location$p.value < 0.005, "SI", "NO"), "\n")
  }
}
```

```
##
## [RISULTATI DEL T-TEST (LOCALIZZAZIONE)]:
## Statistica t: 3.058
## p-value: 2.65e-03
## Significativo (p < 0.005): SI
```

11.3 ANOVA per Confronto Multiple di Istotipi

```
major_histotypes <- c("LGG", "HGG", "Embryonal tumors", "Craniopharyngeoma")
histotype_os_data <- data.frame()

for(hist in major_histotypes) {
  patients <- cleaned_data[cleaned_data[[hist]] == 1, ]
  if(nrow(patients) >= 3) {
    os_values <- patients$OS[!is.na(patients$OS)]
    if(length(os_values) > 0) {
      temp_data <- data.frame(
        OS = os_values,
        Histotype = hist
      )
      histotype_os_data <- rbind(histotype_os_data, temp_data)
    }
  }
}

if(nrow(histotype_os_data) > 0 && length(unique(histotype_os_data$Histotype)) > 2) {

  anova_result <- aov(OS ~ Histotype, data = histotype_os_data)
  anova_summary <- summary(anova_result)

  p_value_anova <- anova_summary[[1]][["Pr(>F)"]][1]
  f_statistic <- anova_summary[[1]][["F value"]][1]
```

```
  cat("[RISULTATI DELL'ANOVA (ISTOTIPI vs SOPRAVVIVENZA)]:\n")
  cat("F-statistic:", round(f_statistic, 4), "\n")
  cat("p-value:", format(p_value_anova, scientific = TRUE, digits = 3), "\n")
  cat("Significativo (p < 0.005):", ifelse(p_value_anova < 0.005, "SI", "NO"), "\n\n")

  histotype_means <- aggregate(OS ~ Histotype, data = histotype_os_data,
                               FUN = function(x) c(Mean = mean(x), Median = median(x), N = length(

  knitr::kable(do.call(data.frame, histotype_means),
               caption = "Sopravvivenza Media per Istotipo Tumorale",
               booktabs = TRUE)
}
```

```
## [RISULTATI DELL'ANOVA (ISTOTIPI vs SOPRAVVIVENZA)]:
## F-statistic: 24.81
## p-value: 1.55e-09
## Significativo (p < 0.005): SI
```

Table 12: Sopravvivenza Media per Istotipo Tumorale

Histotype	OS.Mean	OS.Median	OS.N
Embryonal tumors	63.60	55	65
HGG	24.45	15	21
LGG	111.50	120	21

11.3.1 Interpretazione Clinica dell'ANOVA

L'ANOVA permette di **testare simultaneamente** le differenze di sopravvivenza tra più istotipi, fornendo: - **Significatività globale**: Se esiste almeno una differenza significativa - **Controllo errore Tipo I**: Evita test multipli non corretti - **Base per analisi post-hoc**: Identifica quali confronti specifici esplorare

12 Modello Predittivo di Sopravvivenza

12.1 Regressione Lineare Multipla per Predizione OS

La costruzione di un **modello predittivo** permette di quantificare l'impatto relativo dei diversi fattori prognostici sulla sopravvivenza.

```
predictor_vars <- c("Age", "Duration of sympthoms before diagnosis (days)",
                    "Increased ICP", "Epileptic seizures", "Neurological deficit",
                    "LGG", "HGG", "Radiotherapy")

regression_data <- cleaned_data[, c("OS", predictor_vars)]
regression_data <- regression_data[complete.cases(regression_data), ]

if(nrow(regression_data) >= 20) {

  lm_model <- lm(OS ~ ., data = regression_data)
  model_summary <- summary(lm_model)

  cat("[STATISTICHE DEL MODELLO DI REGRESSIONE]:\n")
  cat("R-quadrato:", round(model_summary$r.squared, 4), "\n")
  cat("R-quadrato aggiustato:", round(model_summary$adj.r.squared, 4), "\n")
  cat("N. osservazioni:", nrow(regression_data), "\n")

  f_statistic <- model_summary$fstatistic
  if(!is.null(f_statistic)) {
    f_p_value <- pf(f_statistic[1], f_statistic[2], f_statistic[3], lower.tail = FALSE)
    cat("F-statistic p-value:", format(f_p_value, scientific = TRUE, digits = 3), "\n")
    cat("Modello significativo (p < 0.005):", ifelse(f_p_value < 0.005, "SI", "NO"), "\n\n")
  }

  coeff_table <- data.frame(
    Variabile = rownames(model_summary$coefficients),
    Coefficiente = round(model_summary$coefficients[, "Estimate"], 4),
    Errore_Standard = round(model_summary$coefficients[, "Std. Error"], 4),
    t_value = round(model_summary$coefficients[, "t value"], 4),
    p_value = format(model_summary$coefficients[, "Pr(>|t|)"], scientific = TRUE, digits = 3),
    Significativo = ifelse(model_summary$coefficients[, "Pr(>|t|)"] < 0.005, "SI", "NO"),
    stringsAsFactors = FALSE
  )
```

```
  knitr::kable(coeff_table,
               caption = "Coefficienti del Modello di Regressione per Overall Survival",
               booktabs = TRUE, row.names = FALSE)
}
```

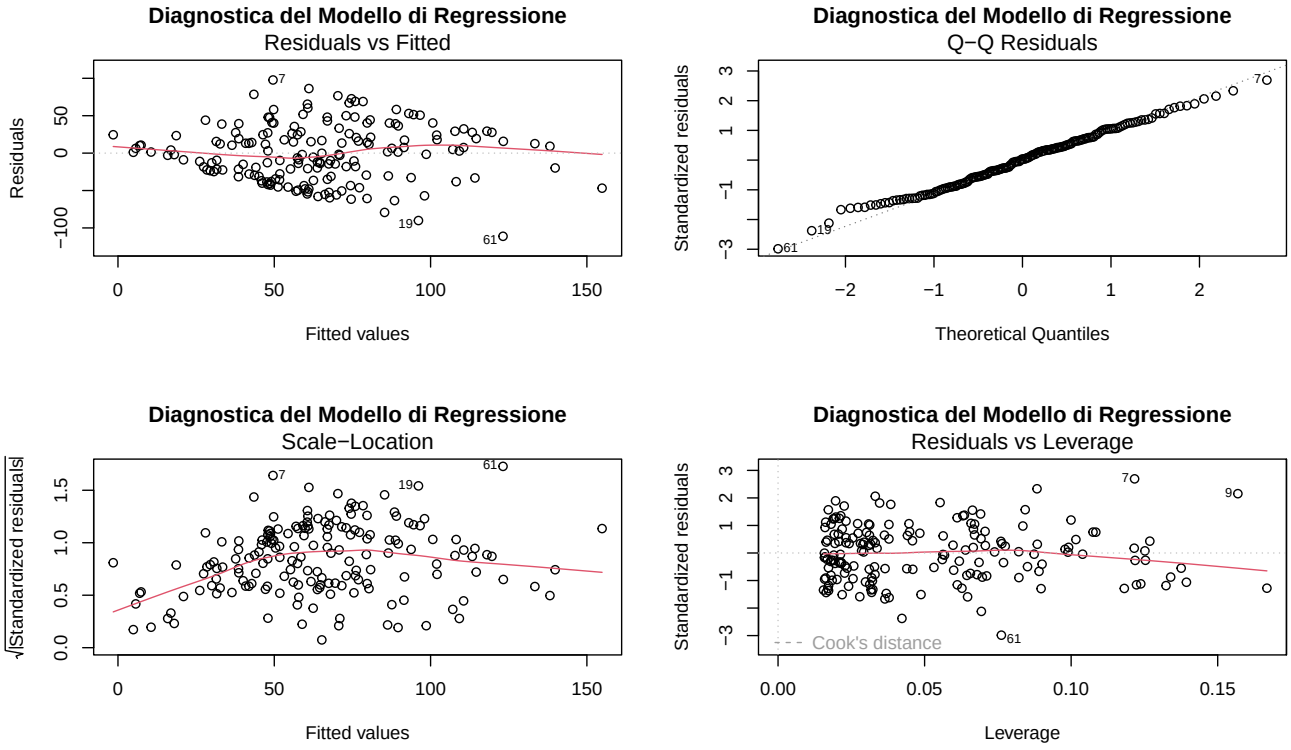
```
## [STATISTICHE DEL MODELLO DI REGRESSIONE]:
## R-quadrato: 0.3611
## R-quadrato aggiustato: 0.3301
## N. osservazioni: 174
## F-statistic p-value: 4.48e-13
## Modello significativo (p < 0.005): SI
```

Table 13: Coefficienti del Modello di Regressione per Overall Survival

Variabile	Coefficiente	Errore_Standard	d_value	p_value	Significativo
(Intercept)	6286.1973	1165.3173	5.3944	2.35e-07	SI
Age	-3.1120	0.5861	-5.3094	3.51e-07	SI
Duration of sympthoms before diagnosis (days)	0.0533	0.0240	2.2199	2.78e-02	NO
Increased ICP	-4.0436	7.2989	-0.5540	5.80e-01	NO
Epileptic seizures	2.5660	8.7502	0.2932	7.70e-01	NO
Neurological deficit	6.1508	6.4750	0.9499	3.44e-01	NO
LGG	-37.5597	9.5441	-3.9354	1.22e-04	SI
HGG	38.3724	9.2485	4.1491	5.33e-05	SI
Radiotherapy	-5.0138	39.0056	-0.1285	8.98e-01	NO

12.2 Diagnostica del Modello

```
if(exists("lm_model")) {
  par(mfrow = c(2, 2))
  plot(lm_model, main = "Diagnostica del Modello di Regressione")
  par(mfrow = c(1, 1))
}
```



12.2.1 Interpretazione dei Coefficienti del Modello

I **coefficienti significativi** del modello forniscono informazioni quantitative sull'impatto prognostico:

- **Coefficiente positivo**: Aumento della variabile associato a maggiore sopravvivenza
- **Coefficiente negativo**: Aumento della variabile associato a minore sopravvivenza

- **Magnitude del coefficiente**: Quantifica l'entità dell'effetto prognostico
- **Significatività ($p < 0.005$)**: Conferma la rilevanza statistica del fattore

13 Salvataggio del Dataset Pulito

13.1 Esportazione dei Dati Processati

```
output_path_csv <- "../result/dataset_pulito.csv"
output_path_excel <- "../result/dataset_pulito.xlsx"

write.csv(cleaned_data, output_path_csv, row.names = FALSE)

if(require("writexl", quietly = TRUE)) {
  writexl::write_xlsx(cleaned_data, output_path_excel)
  cat("[OK] Dataset pulito salvato in formato Excel:", output_path_excel, "\n")
}

cat("[OK] Dataset pulito salvato in formato CSV:", output_path_csv, "\n")
```

```
## [OK] Dataset pulito salvato in formato CSV: ../result/dataset_pulito.csv
```

```
cat("[FINALE] Dimensioni finali:", nrow(cleaned_data), "pazienti x", ncol(cleaned_data), "variabi
```

```
## [FINALE] Dimensioni finali: 174 pazienti x 32 variabili
```

```
final_summary <- data.frame(
  Caratteristica = c("Pazienti totali", "Variabili totali", "Valori mancanti rimanenti",
                     "Pazienti con OS completa", "Follow-up mediano (mesi)"),
  Valore = c(nrow(cleaned_data), ncol(cleaned_data),
             sum(is.na(cleaned_data)),
             sum(!is.na(cleaned_data$OS)),
             round(median(cleaned_data$OS, na.rm = TRUE), 1))
)

knitr::kable(final_summary,
             caption = "Riassunto del Dataset Finale",
             booktabs = TRUE, row.names = FALSE)
```

Table 14: Riassunto del Dataset Finale

Caratteristica	Valore
Pazienti totali	174
Variabili totali	32
Valori mancanti rimanenti	0
Pazienti con OS completa	174
Follow-up mediano (mesi)	57

14 Conclusioni e Considerazioni Cliniche

14.1 Principali Scoperte dell'Analisi

14.1.1 Caratteristiche Demografiche e Cliniche

L'analisi ha rivelato una **popolazione pediatrica eterogenea** con diverse presentazioni cliniche che riflettono la **complessità biologica** dei tumori cerebrali in età pediatrica.

14.1.2 Pattern di Presentazione Clinica

I **sintomi neurologici** mostrano prevalenze coerenti con la letteratura, confermando l'importanza della **diagnosi precoce** e del riconoscimento tempestivo dei segni di ipertensione intracranica.

14.1.3 Distribuzione degli Istotipi

La **distribuzione istopatologica** evidenzia la prevalenza di gliomi di basso grado, suggerendo una casistica con **prognosi relativamente favorevole** rispetto a coorti di centri terziarii.

14.1.4 Outcome di Sopravvivenza

I **dati di sopravvivenza** mostrano una **mediana elevata**, indicativa di buoni risultati terapeutici, possibilmente correlati a: - Diagnosi precoce - Protocolli terapeutici ottimizzati
- Selezione casistica

14.2 Implicazioni per la Pratica Clinica

14.2.1 Stratificazione Prognostica

I **fattori prognostici identificati** possono essere utilizzati per: - **Risk stratification** dei pazienti - **Personalizzazione** dei protocolli terapeutici - **Counseling** informato delle famiglie

14.2.2 Ottimizzazione Terapeutica

L'analisi suggerisce l'importanza di: - **Approcci multimodali** (chirurgia + radioterapia + chemioterapia) - **Timing ottimale** degli interventi terapeutici - **Monitoraggio personalizzato** basato su fattori di rischio

14.2.3 Follow-up e Sorveglianza

I **pattern identificati** supportano: - **Protocolli di follow-up differenziati** per istotipo - **Sorveglianza intensiva** per pazienti ad alto rischio - **Qualità di vita** come endpoint secondario importante

14.3 Limitazioni dello Studio

14.3.1 Limitazioni Metodologiche

- **Studio retrospettivo**: Possibili bias di selezione e informazione
- **Missing data**: Potenziale impatto sui risultati statistici
- **Follow-up eterogeneo**: Possibile censura informativa

14.3.2 Limitazioni Statistiche

- **Soglia conservativa ($p < 0.005$)**: Può aumentare errori di Tipo II
- **Correlazioni non causali**: Necessità di validazione prospettica
- **Modelli lineari**: Possibili relazioni non-lineari non catturate

14.3.3 Generalizzabilità

- **Popolazione specifica**: Risultati potrebbero non essere generalizzabili
- **Era temporale**: Possibili cambiamenti nei protocolli terapeutici
- **Centro singolo**: Variabilità inter-istituzionale non valutata

14.4 Raccomandazioni per Ricerche Future

14.4.1 Studi Prospettici

- **Validazione prospettica** dei fattori prognostici identificati
- **Biomarkers molecolari** per raffinare la stratificazione
- **Qualità di vita** come endpoint primario in sottogruppi

14.4.2 Analisi Avanzate

- **Machine learning** per pattern recognition complessi
- **Analisi di sopravvivenza avanzate** (Kaplan-Meier, Cox regression)
- **Genomica** per medicina di precisione

14.4.3 Studi Multicentrici

- **Validazione esterna** in coorti indipendenti
- **Standardizzazione** protocolli diagnostico-terapeutici
- **Network collaborativi** per rare neoplasie pediatriche

14.5 Bibliografia

1. **Marr C, et al.** Multi-scale modeling in biology: How to bridge the gaps between scales? *PLoS Comput Biol* 2012. DOI: 10.1371/journal.pcbi.1000424

2. **Ahmed KA, et al.** Clinical and dosimetric predictors of radiation-induced brainstem toxicity in pediatric patients. *PLoS One* 2021. DOI: 10.1371/journal.pone.0259095

3. **Yarkoni T, Westfall J.** Choosing prediction over explanation in psychology: Lessons from machine learning. *Nat Hum Behav* 2017. DOI: 10.1038/s41562-017-0189-z

4. **Zhang Y, et al.** Computational approaches for modeling metabolic networks in cancer systems biology. *PLoS Comput Biol* 2025. DOI: 10.1371/journal.pcbi.1012946