Analisi Pulizia Dati e Esplorativa: Effetto Analgesia Epidurale su Progressione Cancro Colon-Retto

Stefano Quartuccio

2025-09-04

Introduzione

Questo report presenta l'analisi di pulizia dati e esplorativa per lo studio retrospettivo sull'effetto dell'analgesia epidurale (EA) sulla progressione del carcinoma del colon-retto in stadio IV dopo resezione del tumore primario.

Obiettivo dello studio: Valutare se l'uso dell'analgesia epidurale durante l'intervento chirurgico influenzi la sopravvivenza e la progressione della malattia nei pazienti con carcinoma del colon-retto metastatico.

Dataset: 999 pazienti operati tra 2014-2017, con follow-up fino a 5 anni.

Processo di Pulizia Dati

Prima di procedere con l'analisi esplorativa, è stato applicato un processo di pulizia avanzata per garantire la qualità dei dati.

Problemi Identificati nel Dataset Grezzo

```
# Carica dataset originale per confronto
data_raw <- readRDS("../data/2018-07-20/dataset_cleaned.rds")
data_clean <- readRDS("../results/dataset_final_cleaned.rds")

# Analisi valori mancanti originali
missing_original <- sapply(data_raw, function(x) sum(is.na(x)))
missing_original <- missing_original[missing_original > 0]

cat("### Valori Mancanti nel Dataset Originale\n")
```

```
## ### Valori Mancanti nel Dataset Originale
```

```
if(length(missing_original) > 0) {
  for(i in 1:length(missing_original)) {
    cat("- **", names(missing_original)[i], "**: ", missing_original[i], " valori mancanti (",
        round(missing_original[i]/nrow(data_raw)*100, 1), "%)\n", sep="")
  }
} else {
```

```
  cat("Nessun valore mancante rilevato.\n")
}
```

```
## - **cea**: 19 valori mancanti (1.9%)
## - **log_cea**: 19 valori mancanti (1.9%)
## - **cell_diff**: 55 valori mancanti (5.5%)
## - **mucin_type**: 57 valori mancanti (5.7%)
## - **signet_ring**: 57 valori mancanti (5.7%)
## - **lymphovascularinvasion**: 54 valori mancanti (5.4%)
## - **perineural**: 57 valori mancanti (5.7%)
```

```
cat("\n### Statistiche Outliers CEA\n")
```

```
##
## ### Statistiche Outliers CEA
```

```
cea_summary <- summary(data_raw$cea)
cat("- **CEA range**: ", round(min(data_raw$cea, na.rm=TRUE), 2), " - ",
    round(max(data_raw$cea, na.rm=TRUE), 2), "\n", sep="")
```

```
## - **CEA range**: 0.44 - 15126
```

```
cat("- **CEA mediana**: ", round(median(data_raw$cea, na.rm=TRUE), 2), "\n", sep="")
```

```
## - **CEA mediana**: 18.9
```

```
cat("- **Casi CEA > 1000**: ", sum(data_raw$cea > 1000, na.rm=TRUE), "\n", sep="")
```

```
## - **Casi CEA > 1000**: 60
```

```
cat("- **Casi CEA > 5000**: ", sum(data_raw$cea > 5000, na.rm=TRUE), "\n", sep="")
```

```
## - **Casi CEA > 5000**: 8
```

Strategie di Pulizia Applicate

1. Gestione Valori Mancanti

- **CEA e log_CEA** (19 NA): Imputazione con mediana per preservare la distribuzione
- **Variabili istologiche** (cell_diff, mucin_type, signet_ring, perineural, lymphovascularinvasion): Creazione categoria "Sconosciuto" (codice 9) invece di eliminare osservazioni

2. Gestione Outliers

- **CEA estremi**: Applicata winsorization al 99° percentile
- **Motivazione**: Valori estremi clinicamente possibili ma che possono distorcere analisi statistiche

3. Correzione Inconsistenze Categoriali

- **RBC**: Valore anomalo "2" corretto a "1" (trasfusione sì)
- **Cell differentiation**: Valore "0" (non valido) corretto a "9" (sconosciuto)
- **Conversione fattori**: Tutte le variabili categoriche convertite a factor per analisi statistiche

Risultati della Pulizia

```
cat("### Confronto Prima/Dopo Pulizia\n")
```

```
## ### Confronto Prima/Dopo Pulizia
```

```
cat("| Metrica | Prima | Dopo | Miglioramento |\n")
```

```
## | Metrica | Prima | Dopo | Miglioramento |
```

```
cat("|---------|-------|------|---------------|\n")
```

```
## |---------|-------|------|---------------|
```

```
cat("| Valori mancanti | ", sum(is.na(data_raw)), " | ", sum(is.na(data_clean)), " |   Eliminati |\n", s
```

```
## | Valori mancanti | 318 | 0 |   Eliminati |
```

```
cat("| CEA massimo | ", round(max(data_raw$cea, na.rm=TRUE), 0), " | ",
    round(max(data_clean$cea, na.rm=TRUE), 0), " |   Winsorizzato |\n", sep="")
```

```
## | CEA massimo | 15126 | 3843 |   Winsorizzato |
```

```
cat("| Variabili factor | ", sum(sapply(data_raw, is.factor)), " | ",
    sum(sapply(data_clean, is.factor)), " |   Ottimizzate |\n", sep="")
```

```
## | Variabili factor | 0 | 23 |   Ottimizzate |
```

```
cat("\n### Dataset Finale\n")
```

```
##
## ### Dataset Finale
```

```
cat("- **Dimensioni**: ", nrow(data_clean), " osservazioni × ", ncol(data_clean), " variabili\n", sep="
```

```
## - **Dimensioni**: 999 osservazioni × 32 variabili
```

```
cat("- **Qualità**: Zero valori mancanti, codifica consistente, outliers gestiti\n")
```

```
## - **Qualità**: Zero valori mancanti, codifica consistente, outliers gestiti
```

```
cat("- **Pronto per analisi statistiche avanzate**\n")
```

```
## - **Pronto per analisi statistiche avanzate**
```

Caricamento e Descrizione Dataset

```
# Carica il dataset FINALMENTE pulito
# Il dataset è stato sottoposto a pulizia avanzata: valori mancanti gestiti,
# outliers winsorizzati, inconsistenze corrette, tipi di dati ottimizzati
data <- readRDS("../results/dataset_final_cleaned.rds")

# Dimensioni del dataset finale
cat("Dataset finale dimensioni:", nrow(data), "osservazioni ×", ncol(data), "variabili\n\n")
```

```
## Dataset finale dimensioni: 999 osservazioni × 32 variabili
```

```
# Verifica qualità dati finale
cat("### Qualità Dataset Finale\n")
```

```
## ### Qualità Dataset Finale
```

```
cat("- Valori mancanti totali:", sum(is.na(data)), "\n")
```

```
## - Valori mancanti totali: 0
```

```
cat("- Variabili factor:", sum(sapply(data, is.factor)), "\n")
```

```
## - Variabili factor: 23
```

```
cat("- Variabili numeriche:", sum(sapply(data, is.numeric)), "\n\n")
```

```
## - Variabili numeriche: 9
```

```
# Mostra le prime righe per overview
head(data)
```

```
## # A tibble: 6 x 32
##   bian_ma   age gender asa   asa3  dm    cad   hf    cva   ckd       cea log_cea
##     <dbl> <dbl> <fct>  <fct> <fct> <fct> <fct> <fct> <fct> <fct>   <dbl>   <dbl>
## 1       2    52 1      3     1     0     0     0     0     0      937      6.84
## 2      21    85 2      2     0     0     1     0     0     0      134      4.90
## 3      22    45 2      2     0     0     0     0     0     0     1142      7.04
## 4      23    57 2      2     0     0     0     0     0     0       31.4    3.45
## 5      29    43 1      2     0     1     0     0     0     0        2.31   0.837
## 6      33    51 2      2     0     0     0     0     0     0      262      5.57
## # i 20 more variables: laparoscopic <fct>, tumor_loc <dbl>, ea <fct>,
## #   anes_time <dbl>, log2at <dbl>, rbc <fct>, ajcc <fct>, liver_only <fct>,
## #   cell_diff <fct>, mucin_type <fct>, signet_ring <fct>,
## #   lymphovascularinvasion <fct>, perineural <fct>, ct <fct>, rt <fct>,
## #   nactrt <fct>, death <fct>, interval <dbl>, progress <fct>, interval_r <dbl>
```

Descrizione Variabili Principali

Variabili Demografiche

- **age**: Età del paziente in anni
- **gender**: Genere (1 = maschio, 2 = femmina)

Comorbidità e Stato di Salute

- **asa**: Score ASA (American Society of Anesthesiologists) - classificazione stato fisico:
  - 1: Paziente sano
  - 2: Malattia sistemica lieve
  - 3: Malattia sistemica grave
  - 4: Malattia sistemica grave che minaccia la vita
- **asa3**: Indicatore binario (1 se ASA 3, 0 altrimenti)
- **dm**: Diabete mellito (1 = presente, 0 = assente)
- **cad**: Malattia coronarica (1 = presente, 0 = assente)
- **hf**: Insufficienza cardiaca (1 = presente, 0 = assente)
- **cva**: Ictus cerebrale (1 = presente, 0 = assente)
- **ckd**: Malattia renale cronica (1 = presente, 0 = assente)

Variabili Oncologiche

- **cea**: Livello di antigene carcinoembrionale (CEA) - marker tumorale
- **log_cea**: Logaritmo del CEA (trasformazione per normalizzare distribuzione skewed)
- **ajcc**: Stadio secondo classificazione AJCC (American Joint Committee on Cancer):
  - 4a: Metastasi limitate a un organo
  - 4b: Metastasi multiple organi
- **liver_only**: Metastasi esclusivamente epatica (1 = sì, 0 = no)
- **cell_diff**: Grado di differenziazione cellulare (1 = ben differenziato, 2 = moderatamente, 3 = scarsamente)
- **mucin_type**: Carcinoma mucinoso (1 = sì, 0 = no)
- **signet_ring**: Cellule a anello con castone (1 = sì, 0 = no)
- **lymphovascularinvasion**: Invasione linfonodale/vascolare (1 = presente, 0 = assente)
- **perineural**: Invasione perineurale (1 = presente, 0 = assente)

Variabili di Trattamento

- **laparoscopic**: Intervento laparoscopico (1 = sì, 0 = no)
- **ea**: Analgesia epidurale (VARIABILE PRINCIPALE DI INTERESSE - 1 = usata, 0 = no)
- **anes_time**: Durata anestesia in minuti
- **log2at**: Log2 della durata anestesia (trasformazione per analisi)
- **rbc**: Trasfusione globuli rossi (1 = sì, 0 = no)
- **ct**: Chemioterapia (1 = sì, 0 = no)
- **rt**: Radioterapia (1 = sì, 0 = no)
- **nactrt**: Chemioterapia/radioterapia neoadiuvante (1 = sì, 0 = no)

Outcome

- **death**: Morte (1 = deceduto, 0 = vivo)
- **progress**: Progressione malattia (1 = progressione, 0 = no)
- **interval**: Tempo in mesi fino all'evento (morte o progressione)
- **interval_r**: Tempo ricodificato (probabilmente per analisi di sopravvivenza con dati censurati)

Analisi Esplorativa

```
# Statistiche descrittive di base
summary(data)
```

```
##     bian_ma          age        gender  asa     asa3    dm      cad     hf
##  Min.   :   2   Min.   :18.00   1:612   1: 53   0:612   0:796   0:927   0:955
##  1st Qu.:1270   1st Qu.:55.00   2:387   2:559   1:387   1:203   1: 72   1: 44
##  Median :2427   Median :65.00           3:366
##  Mean   :2522   Mean   :65.18           4: 20
##  3rd Qu.:3885   3rd Qu.:77.00           5:  1
##  Max.   :5172   Max.   :98.00
##  cva     ckd          cea              log_cea       laparoscopic
##  0:942   0:861   Min.   :   0.44   Min.   :-0.821   0:961
##  1: 57   1:138   1st Qu.:   4.09   1st Qu.: 1.409   1: 38
##                  Median :  18.89   Median : 2.939
##                  Mean   : 212.63   Mean   : 3.191
##                  3rd Qu.:  86.17   3rd Qu.: 4.456
##                  Max.   :3843.40   Max.   : 8.254
##    tumor_loc      ea        anes_time         log2at      rbc     ajcc
##  Min.   :0.0000   0:834   Min.   : 45.0   Min.   :5.492   0:577   4a:558
##  1st Qu.:0.0000   1:165   1st Qu.:255.0   1st Qu.:7.994   1:422   4b:441
##  Median :0.0000           Median :315.0   Median :8.299
##  Mean   :0.3133           Mean   :338.4   Mean   :8.313
##  3rd Qu.:1.0000           3rd Qu.:390.0   3rd Qu.:8.607
##  Max.   :1.0000           Max.   :960.0   Max.   :9.907
##  liver_only cell_diff mucin_type signet_ring lymphovascularinvasion perineural
##  0:629      1:813     0:869      0:900       0:460                  0:734
##  1:370      2:121     1: 73      1: 42       1:485                  1:208
##             9: 65     9: 57      9: 57       9: 54                  9: 57
##
##
##
##  ct      rt      nactrt  death      interval         progress
##  0:110   0:889   0:844   0:572   Min.   :  0.03285   0:221
##  1:889   1:110   1:155   1:427   1st Qu.:  7.34292   1:778
##                                  Median : 17.47844
##                                  Mean   : 24.22519
##                                  3rd Qu.: 31.73717
##                                  Max.   :135.78645
##    interval_r
##  Min.   :  0.03285
##  1st Qu.:  2.95688
##  Median :  5.74949
```

```
##  Mean   : 11.26982
##  3rd Qu.: 12.59959
##  Max.   :134.20945
```

```r
# Valori mancanti per colonna
missing_summary <- sapply(data, function(x) sum(is.na(x)))
missing_summary <- missing_summary[missing_summary > 0]
if(length(missing_summary) > 0) {
  cat("\nValori mancanti per colonna:\n")
  print(missing_summary)
} else {
  cat("\nNessun valore mancante nel dataset.\n")
}
```

```
##
## Nessun valore mancante nel dataset.
```

Analisi Variabile Principale: Analgesia Epidurale

```r
# Distribuzione dell'analgesia epidurale
ea_table <- table(data$ea)
ea_prop <- prop.table(ea_table)

cat("Distribuzione Analgesia Epidurale:\n")
```

```
## Distribuzione Analgesia Epidurale:
```

```r
cat("No EA:", ea_table["0"], "pazienti (", round(ea_prop["0"] * 100, 1), "%)\n")
```

```
## No EA: 834 pazienti ( 83.5 %)
```

```r
cat("Con EA:", ea_table["1"], "pazienti (", round(ea_prop["1"] * 100, 1), "%)\n\n")
```

```
## Con EA: 165 pazienti ( 16.5 %)
```

```r
# Confronto outcome tra gruppi
death_by_ea <- table(data$ea, data$death)
death_rates <- prop.table(death_by_ea, margin = 1)

cat("Tasso di mortalità per gruppo:\n")
```

```
## Tasso di mortalità per gruppo:
```

```r
cat("Senza EA:", round(death_rates["0", "1"] * 100, 1), "%\n")
```

```
## Senza EA: 43.2 %
```

```r
cat("Con EA:", round(death_rates["1", "1"] * 100, 1), "%\n")
```

```
## Con EA: 40.6 %
```

Analisi Demografica

```r
# Età per genere
cat("Età media per genere:\n")
```

```
## Età media per genere:
```

```r
tapply(data$age, data$gender, mean, na.rm = TRUE)
```

```
##        1        2
## 66.42974 63.19897
```

```r
# Distribuzione ASA score
asa_table <- table(data$asa)
cat("\nDistribuzione ASA score:\n")
```

```
##
## Distribuzione ASA score:
```

```r
print(asa_table)
```

```
##
##   1   2   3   4   5
##  53 559 366  20   1
```

Analisi Comorbidità

```r
# Prevalenza comorbidità
comorbidities <- c("dm", "cad", "hf", "cva", "ckd")
comorb_prev <- sapply(data[, comorbidities], function(x) mean(as.numeric(as.character(x)), na.rm = TRUE)

cat("Prevalenza comorbidità (%):\n")
```

```
## Prevalenza comorbidità (%):
```

```r
print(round(comorb_prev, 1))
```

```
##   dm  cad   hf  cva  ckd
## 20.3  7.2  4.4  5.7 13.8
```

Conclusioni Preliminari

Da questa analisi esplorativa iniziale:

1. **Campione**: 999 pazienti con carcinoma colon-retto stadio IV
2. **Esposizione**: Circa 16.5% dei pazienti ha ricevuto analgesia epidurale
3. **Outcome**: Tasso di mortalità del 42.7%
4. **Follow-up**: Tempo medio di osservazione di 24.2 mesi

Prossimi passi dell'analisi: - Analisi di sopravvivenza (Kaplan-Meier, Cox regression) - Regressione logistica per identificare fattori prognostici - Analisi di sensitività per confounding - Valutazione dell'effetto dell'EA stratificato per sottogruppi

Questa analisi richiede tecniche statistiche più avanzate per rispondere alla domanda di ricerca principale sull'impatto dell'analgesia epidurale sulla progressione tumorale.