HOLMUSK

Data Challenge

Javier Manzano

Agenda

- Problem Statement & Databases
  - “Event_duration.csv” + “Patient_characteristics.csv”
- Data Wrangling Process
  - Merging, Cleaning, …
- Inconsistencies
  - “treatment_variable”
- Survival Analysis
  - Approach + Methods
  - Results + Insigths

Problem Statement & Databases

- Problem Statement (some points to consider):
  - Patient diagnosed with a specific condition and prescribed either **Drug A** or **Drug B** for the treatment
  - The patients are monitored for the **occurrence of a specific event after the start of the treatment**
  - To compare the real-world **efficacy of the two drugs** by comparing the risk of events: survival analysis
    - The two groups **may not have balanced patient characteristics**
    - How to measure and reduce the impact of this, in the analysis approach

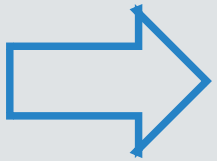
- 2 databases:
  - "Event_duration.csv" -> [**ed**] -> 19284 observations and 4 columns
  - "Patient_characteristics.csv" -> [**pc**] -> 19284 observations and 37 columns

Merging and Cleaning Processes

- Merging: [ed] & [pc]
  - Inner Join: key field -> "**patient_id**“
    - 15868 observations
    - [ed] and [pc] -> “treatmentVariable_ed” and “treatmentVariable_pc”

patient_id
0
1
2
3

treatmentVariable_ed	treatmentVariable_pc
Drug_A	Drug_A
Drug_A	Drug_A
Drug_A	Drug_A
Drug_A	Drug_A
Drug_B	Drug_B



treatmentVariable_ed	treatmentVariable_pc
Drug_A	Drug_A
Drug_A	Drug_A
Drug_A	Drug_A
Drug_A	Drug_A
Drug_B	Drug_B
Drug_B	Drug_A
Drug_A	Drug_B
Drug_A	Drug_B
Drug_B	Drug_A

Inconsistencies

- From 15868 observations
  - 8057 observations (~51%): “treatmentVariable_ed” = “treatmentVariable_pc”
  - 7811 observations (~49%): “treatmentVariable_ed” != “treatmentVariable_pc”

- The same patient could not have taken both of the drugs at the same time
  - It may not be clear to identify the effect of either drug A or drug B
  - 7811 observations (~49%) with inconsistencies

- Consequently
  - I decided to select observations w/o inconsistencies
  - 8057 observations (~51%)

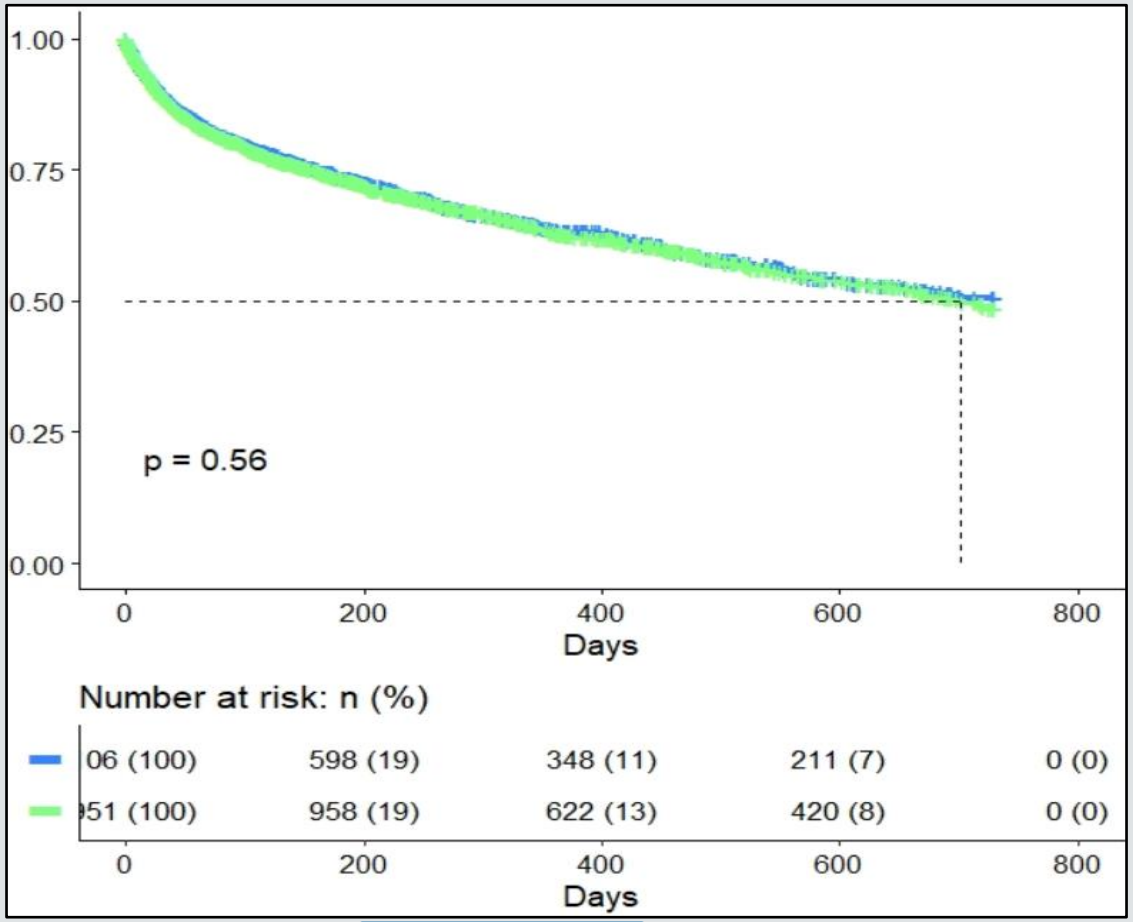
Survival Analysis

- Approach
  - 8057 observations with **unbalanced** patient characteristics
  - Processing from years to months and **days** (“yearsInDays”)
  - Categorizing age variable (median) in “ageCat”: adult (<= 79) & senior (80+)
  - W/O lab_2 to lab_8: NA´s
  - SA 1: all the data (n = 8057)
  - Stratified random sampling
    - SA 2: bleedingEvent (n = 3138)
    - SA 3: (+) treatmentVariable (n = 6212)
    - SA 4: (+) sex (n = 6640)
    - SA 5: (+) age -> ageCat (n = 7504)

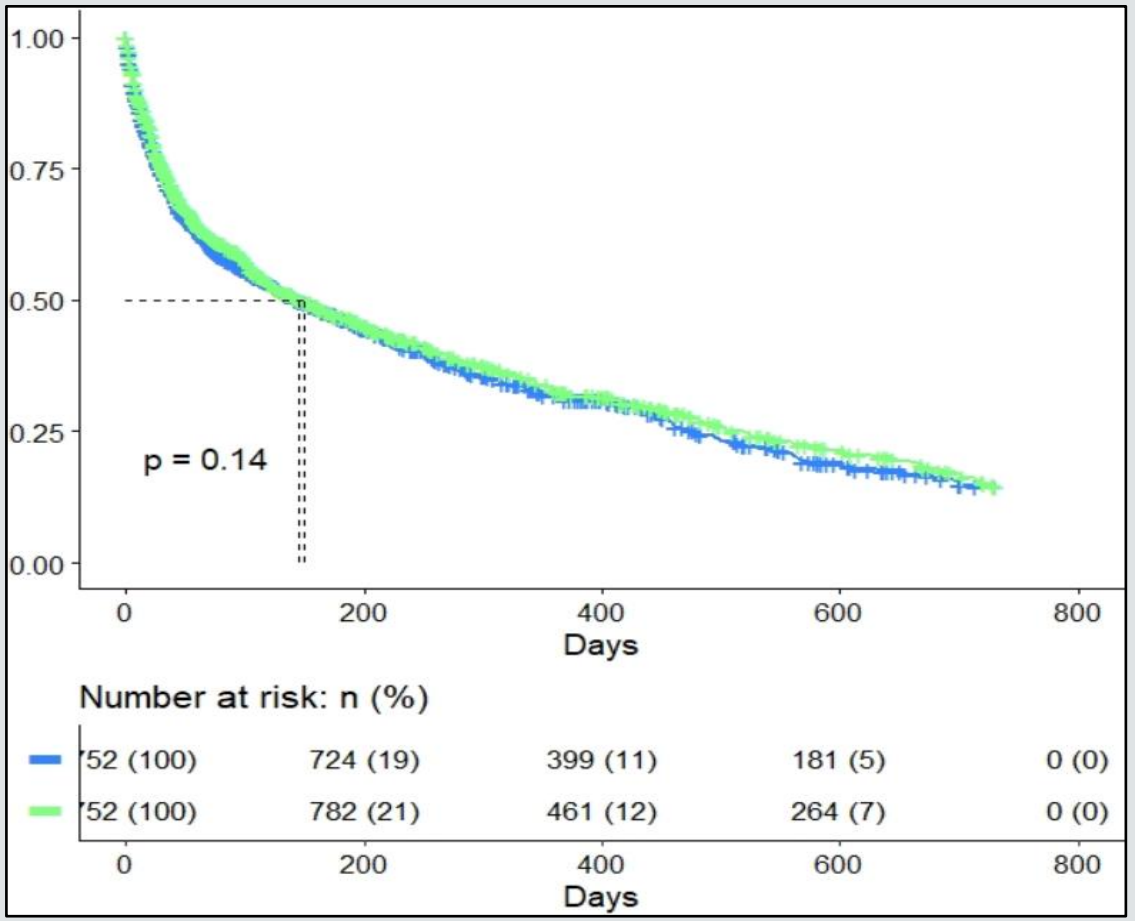
- Methods
  - Kaplan-Meier Method
  - Cox Method (1+ variables)

Survival Analysis

- Results (I)*: Survival Curves between SA 1 and SA 5



SA 1 | Drug A | Drug B



SA 5 | Drug A | Drug B

*Check the *"resultsSummary_FJMM.pdf"* and *"code_dataChallenge_FJMM.R"* files (in GitHub) for more details

Survival Analysis

- Results (II)*: Kapla-Meier Method and Cox Method

SA	Method	
	Kaplan-Meier (Log-Rank test)	**Cox** (Hazard ratios)
SA 1 (n = 8057)	There is no statistically significant difference between Drug A and Drug B	Variables (+) sex; age/ageCat; other_drugs_1to8; diagnosis_1to15; lab_1; Diag_Score_1to2
		Statistically significant at 5% None
SA 5 (n = 7504)	There is no statistically significant difference between Drug A and Drug B	Variables (+) sex; age/ageCat; other_drugs_1to8; diagnosis_1to15; lab_1; Diag_Score_1to2
		Statistically significant at 5% treatmentVariableDrug; other_drugs_5; other_drugs_8; diagnosis_4

*Check the *"resultsSummary_FJMM.pdf"* and *"code_dataChallenge_FJMM.R"* files (in GitHub) for more details

Survival Analysis

- Insigths (I)
  - The databases presented inconsistencies between the relevant variable records
    - “treatmentVariable”: ed != pc
  - Smaller database, but no inconsistencies
    - To compare the efficacy of the two drugs (A and B)
  - To measure and reduce the impact of the unbalanced patient characteristics
    - Stratified random sampling: bleedingEvent, treatmentVariable, sex and ageCat
    - Balanced sample without losing so many observations
    - n “original” (8057) – n “stratified random sample” (7504) = 553 observations

Survival Analysis

- Insigths (II)
  - Kapla-Meier Method
    - yearsInDays, bleedingEvent and treatmentVariable
    - No differences were identified between the "original" sample (SA 1) and stratified random sample (SA 5)
    - In both, there was no statistically significant difference between drug A and drug B
  - Cox Method (1+ variables)
    - yearsInDays, bleedingEvent, treatmentVariable, sex, age/ageCat, other_drugs_1to8, diagnosis_1to15, lab_1 and Diag_Score_1to2
    - SA 1: None of the variables showed statistical significance at 5%
    - SA 5: treatmentVariableDrug; other_drugs_5; other_drugs_8; diagnosis_4 were statistically significant at 5%
      - Stratified Random Sample (to replicate): it is necessary to use a seed (weakness)

HOLMUSK

Data Challenge

Javier Manzano