Data Analyst R Exam

Introduction:

You are a new data analyst working on a dataset pulled from pediatric proton radiotherapy clinical trial at MGH. The dataset, provided to you as a csv and named "da_exam_file" contains longitudinal information on cancer patients including: When they were consented to the study, when they received radiation treatment, when their treatment ended, and follow-up information. Information collected at follow-up contains dates and statuses on death, recurrence, and presence of secondary tumors.

The following codebook is available to you:

ID: patient study ID

event: Event of study in the following chronological order:

1. Pretreatment
2. Treatment
3. Follow-up (ranges from 1-10 possible)

followup_date: Date of follow-up

consent: Date of enrollment to the trial

radiation_start_date: Proton radiotherapy start date

radiation_end_date: Proton radiotherapy end date

death_date: Date of death

recurrence_date: Date of recurrence

secondary_tumor_date: Date of secondary tumor

death:

0 = No

1 = Yes

recurrence:

0 = No

1 = Yes

secondary_tumor:

0 = No

1 = Yes

Note About Data Collection:

1. The following variables are strictly collected during the follow-up event timepoints:
   a. Death / date of death
   b. Recurrence / date of recurrence
   c. Secondary tumor / date of secondary tumor
2. The Following variables are strictly collected during the treatment timepoints:
   a. Radiation start and end dates
3. Not all patients will have follow-up dates yet as this is an ongoing study.
4. For the sake of this exam, if a patient died, recurred, or had a secondary tumor, there **will** be a corresponding date collected (no missing date data). For example, if death = 1 for a patient, that patient **will have a death date** in this dataset.
5. Not all outcomes (deaths, recurrences, or secondary tumors) will occur at the same follow-up timepoint.

Please see the next page for the exam.

TASKS:

1. The PI wants to be able to view all this data in a simpler format. Specifically, they would like all data to be displayed on one row per patient and written to a csv file. Please convert this long data format to wide and attach the file in your response. Please use the following guideline for a list of data to be represented in wide format and in the following column order:

    a. Desired Order: ID, consent date, date of last follow-up, radiation start and end dates, date of death, date of recurrence, date of secondary tumor, death, recurrence, secondary tumor
    b. For clarity, please **rename** the "date of last follow-up" date variable to "latest_fup_date" within your wide dataset.

    Note: Be careful about merging data, make sure to replace any NA or blanks that occur during your merges within the following variables: **death, recurrence, secondary_tumor**. To be clear, if a patient has follow-up data available and did not die, the final wide dataset should indicate **0** for death for this patient, not NA (same for recurrence and secondary tumor status). If a patient had no follow-up data collected, their status on these variables should be left blank or NA.

    Hint on Merging: The PI would like to see a row for every patient who has a consent date within this dataset, regardless of if they have available follow-up. This way, they can clearly see who the total number of patients is, as well as of those, who is missing follow-up data.

2. In order to prepare the data for publication, the PI would like to know the median follow-up in years we have available in this cohort. For all studies, follow-up duration is calculated as the difference between the last date of follow-up for a patient and the radiation start date. Median follow-up should be calculated only among those who have a follow-up timepoint recorded. To convert to years, please use 365.25

    Median Follow-up (min-max):

3. The PI would like summaries on the following:
    a. Number of patients deceased
    b. Number of patients with recurrences
    c. Number of patients who experienced a secondary tumor
    d. Number of patients that do not have at least 1 follow-up event recorded