ASSIGNMENT 2024

Write your name here:

Akson Sam Varghese

Instructions:

1. *The assignment has 6 questions totalling 55 marks.*
2. *Please note that the questions DO NOT carry equal marks.*
3. *Please attempt all questions.*
4. *Write your answers in this document after each question.*
5. *Please make sure that your code is well-formatted, preferably using a single-spaced font such as Consolas, Courier New, Lucida Console etc.*
6. *Please adhere to all specified word limits.*
7. *Complete the PMIM602 cover-sheet before submitting (next page).*

DEADLINE for submission of all assessment materials:

1pm on Tuesday 28th May 2024

Submit materials for PMIM602 - Advanced Analysis of Linked Health Data through Canvas, Swansea University

Swansea University
Prifysgol Abertawe
Medical School
Ysgol Feddygaeth

Assignment for PMIM602

Module number:	PMIM602
Module name:	Advanced Analysis of Linked Health Data
Title of assignment:	Take home questions
Student ID number:	2311233
Word count:	n/a
Declaration:	I understand the following conditions which apply throughout this course: 1. I confirm that I am the sole author of this work. 2. I understand that proof reading by a third party is discouraged, but if used, records should be available as per guidelines. 3. I understand the need for academic integrity and that all my submitted work will adhere to its principles. 4. I understand that the teaching team will take measures to deter, detect and report any academic misconduct. 5. I agree to my work being submitted to the TurnItIn academic database. 6. I understand the importance of assignment deadlines and the need to seek help in good time where personal circumstances interrupt my work.
Please copy and paste this declaration onto the front of the submission.	

1. A complex data file group comprises a number of linkable data collections including multiple Type 3 datasets and where no individual dataset provides a census of the clinical population of interest for the study. Providing examples, outline the challenges that working with such a file group presents above that which may be experienced using a simple file group of linked datasets. **(6 marks)**

Solution:
Consider an example to internally and externally link and analyse a complex file group for a population of diabetes mellitus health outcome. The four datasets used could be listed as

1. Claims data
2. Pharmaceutical data
3. Death register data
4. Hospital medical records data

The problems faced during working with these files are:

A. **Require data management skill**: Typically claims data, pharma data and hospital medical records contains multiple entries of data for an individual. A new set of skills is required to handle the big data while looking for sparse clinical content across all the files. Following are the list of skills required for managing the big data.
   a. Efficient data search approach: The demand is to work effectively and efficiently with very large number of records per individual. The lack of a census file of all cases of a target condition or service in any one file, requires the analyst to search across the above-mentioned files to identify all eligible individuals for diabetes mellitus. Characterisation of the index records would be difficult as the records would be in different files.
   b. Efficient data merge and storage: Once the data is understood and important columns are identified, the data needs to be merged into 1 file for analysis. Merging the files have a complexity of its own, as the merged file would be large in size. The large files would demand higher computer processing and memory needs. So efficient data merge and storage is required to process the files. Typically claims data and hospital records have multiple entries of an individual.
   c. Data check and cleaning: It is important to perform data check and cleaning multiple files and the respective columns that are used for merging the files. Mismatch of the files could occur due to minor or major problems associated with the complete or partial identifiers.

B. **Information deficit**: In the medical service and pharma data, the information on diagnoses of an individual is not explicitly provided. This could lead to following issues.
   a. Programmatic and data Complexity: When information is not provided there is a need to logically derive the necessary information from the files. This adds to the programmatic complexity for implementation and execution of the logic to link the files. This also adds to the data complexity where the created files are complex to understand and manage.
   b. Knowledge Constraints: It is not possible to understand and learn everything about the data given any time constraint of the project. Linkage of data becomes harder if the given data information of the files is not understood. Thus, multiple files could be more difficult to link and analyse.

C. **Error Propagation:** Compiled data files are error prone that possibly lead to logical or systematic errors which further could be propagated to the final analysis. This could lead to errors that are identified when the actual outcomes do not match the expected outcomes.

D. **Extensive forethought:** In complex file groups, it is mandatory to have a clear forethought about the file processing steps that needs to be performed. The steps could be listed as follows:
   a. Tagging: The records need to be tagged as per the health outcome of interest. For example, in the hospital records, all the records are tagged (set to 1) where the diagnosis codes are matching in the columns.

b. Sequence variables: There is a need to create sequence variable (file sequence / conditional sequence / morbidity sequence) to select only relevant records required for the analysis. For example, assigning conditional sequence variable where records are tagged as diabetes = 1.
c. Cutdown: Filter and subset the dataset only to contain records of interest. The cutting down or shrinking procedure depends on the sequence variables. For example, select only those records where conditional sequence variable = 1.
d. Antepenultimate: Merge the files i.e. row bind the files to create a single file called as antepenultimate secondary platform file. Create new file sequence variables to re-establish the record ordinal sequence.
e. Penultimate: Using the antepenultimate file, create the penultimate file. The file contains selection of those records that has one sequence variable = 1. Create new sequence variables.
f. Ultimate: Using the penultimate file, create the ultimate file. The file contains only those records where morbidity sequence = 1.

2. A number of models are available to partition person-time to account for changing states of a disease in longitudinal linked data research. Provide an overview of three of these models outlining the advantages and disadvantages of each. **(9 marks)**

Solution:

It is important to consider the people involved in the study are from the right time period in a study. This could be achieved by breaking up / partitioning people time as per different states or severity levels of the health condition. There are two types of population, they are:

a. Closed: In this type, the population is of fixed nature. The population is defined and selected based on a specific event. Thus, no new members are included once the study is started or in progress. For example, a study on population of atomic bomb survivors, born in 1980 in Perth. The population is lost only through death or migration. It is identified that over time as the study progresses the selected cohort progressively shrinks.

b. Open: In this type, the population is of dynamic nature. The population is defined and selected based on criteria; for example, geolocation. Thus, new members could be added through birth and immigration even after the study is started or while the study is in progress. The population is lost only through death and emigration. For example, population in Wales.

There are multiple models based on the types of population that approximates effect measures like prevalence of a disease with respect to the person-time. Three of the models from them are:

1. One Stage models
   a. Overview:
      i. This type of model assumes that the type of population is a closed system.
      ii. It is used in scenarios where tracking, follow-up of the patient's health state changes in the population is not possible to be identified at a granular level.
      iii. It is a simple, regional model that was devised during the times when epidemiology was still developing but, this model is applicable till date in the epidemiology domain.
      iv. How it works?
         1. First, obtain a count of people living in a region.
         2. To calculate the effect measures, compare with the health records or death records from that region.
         3. For example: Consider a simple case scenario to calculate the disease prevalence using the one stage model where a dynamic population in a closed system is partitioned into non disease and disease status groups. Then, the population is considered to be in a state of dynamic equilibrium, when the flow of non-diseased population into the diseased state equals the flow of diseased population back into the non-diseased state. Then the prevalence is calculated as

$$p = IR.D/(1 + IR.D)$$

$where,$

$p - prevalence$

IR – incidence rate

D – duration or person time

      v. The model is applicable in cases where the infectious diseases like influenza is to be modelled on a large scale.
   b. Advantage:
      i. One of the main advantages of the model is its simplicity. It is simple model that approximates the effect measures in a research study. Simple model often translates to easier interpretability as well.

    ii. This model works well with the assumption that the type of population is closed system.
    iii. It could be used in scenarios where the follow-up data related to the population is relatively hard to obtain.
    iv. Model implementation cost is cheaper and affordable as the data required for the model is less.

  c. Disadvantage:
    i. The model expects populations in research to be closed system, but in research closed systems are rare.
    ii. Approximation has higher error margin in comparison to other models.
    iii. Approximation is limited to only two states of the health, i.e. Disease and Non-disease. But there could be more states of health, based on disease severity as well. Thus, it cannot be used for modelling information at the granular level.

2. Two Stage models
  a. Overview:
    i. This type of model is an extension to the one stage model.
    ii. The model considers and covers the situation where the diseased group is further partitioned according to the disease severity. For example, the disease severity of a specific disease could be classified as non-disease, mild disease, severe disease.
    iii. The model is capable of modelling effect measures with more precision than one stage model but less than other models like continuous disease severity model.
    iv. How it works?
      1. First, obtain the count of the population in the region exposed to a specific disease.
      2. Split and measure the exposure at severity levels of the disease and calculate the effect measures of interest at these levels.
      3. For example, consider a scenario to calculate disease prevalence. Assume the disease has 3 states, non-disease, mild and severe. Then the, prevalence of the disease for mild disease is calculated as,

$$p_{mild} = D_{mild}(p_{no-disease} IR_{no-disease} + p_{severe} RemR_{severe})$$

$$where,$$

$$IR - incidence\ rate\ of\ no\ disease$$

RemR – remission rate at the severe level

    v. The model expects the type of population to be of a closed system and has attained dynamic equilibrium.

  b. Advantage:
    i. In this model it is possible to approximate better than the one stage model.
    ii. This model accounts for higher granularity in comparison to one stage model.
    iii. This model could be used in the scenario where the follow up data of patients is not possible but still the severity of the disease exposure could be measured.

  c. Disadvantage:
    i. This model is costlier than the one stage model as the data required for the approximation is more granular.
    ii. Even though the calculated effect measures are at more granular than the one stage model, it does not reduce the error margins in comparison to other models.
    iii. The model relies on steady state flow, and expects the data to have the type of population to be closed system which is not favourable in a real-life scenario.

3. Continuous Disease Severity models
    a. Overview:
        i. This model gives the ability to look at the time traces individually because of the availability of different points of data.
        ii. This model approximates closer to reality in comparison to the above mentioned two models.
        iii. It is possible to approximate the effect measures at the individual level, different patient progression from lower to higher levels and the rate of patient progression.
        iv. How it works?
            1. Numerous observations made on individuals is monitored, recorded, and obtained.
            2. The observations are then mapped to the disease severity and traces are drawn at individual level by connecting the disease severity events in an ascending order of the time / period when the study was observed.
            3. The traces forms curves that represent the journey of that individual over the time.
    b. Advantage:
        i. The model approximates reality closer than the one stage and two stage model.
        ii. The model is capable to trace and follow at individual level that enables the system to monitor and provide more precise care.
        iii. The conceptualisation of measuring and calculating effect measures associated to patients at different levels, with different progressions. It is helpful in modelling and the analysis of linked health data where the population dynamics may be built up from numerous observations made on individuals.
    c. Disadvantage:
        i. The model expects a certain order of observable event. It assumes that a certain health event must happen first for the trace of an individual to start. It is possible that an individual can have two events at the same time. The model considers the first event as per occurrence but the second event following the first could be misleading. For example, an individual exposed to disease, opts for the first observable event – ambulatory care, and the immediate next observable event is of the individual visiting the GP, then it misleads the information that, the individual is coming back for care, and is considered as a second episode of care which might not be the case.
        ii. There is no measurement available between two observable events. The only information the traces provide is that at a given point of time the health state of an individual. If the curve trace is exponential between two observable events, the information of why that happened is not available. Thus, the model is limited to the observation events with respect to the time.
        iii. The true trajectory is not represented accurately by the individual traces. The trajectory of individual traces is relative to the starting point of the event that could be random.
        iv. The observable events are not only limited to entry and exit events but transition events as well. This model does not account for transition events that provide more information about the health state of an individual.

3. Incidence density sampling is a method that can be used when selecting a comparison/control group from administrative data in longitudinal research designs. Describe the process of incidence sampling in linked administrative data and provide an example of where it has advantages over other forms of sampling. **(5 marks)**

Solution:

Overview**:**
There are four different forms and purposes of sampling within administrative health data files. They are Random Sampling, Matched Sampling, Incidence Density Sampling, and combinations of the earlier mentioned sampling types.

A case-control study nested within a cohort is a highly efficient method for epidemiology investigation. Incidence density sampling is suitable for a nested case-control study. In incidence density sampling, controls and cases are selected as risk sets. A risk set consists of 1 case and n controls. The controls in risk set are selected such that, the members of the cohort are still at risk at the particular time of occurrence of their corresponding case. In such situations, a case might be a control in previous risk set, but it would be a case in its own risk set, or it might be risk or control at another same time of occurrences those occur afterwards. It is possible that the selected individual could be a control in multiple risk sets. It is not possible to restrict the individual for control selection, if done so then it would lead to a biased estimation.

Process:
Following is the process of implementing incidence density sampling.

1. **Step 1: Sorting the records**: Sort the file for all cases in the descending order of total length of follow up time. The follow up time could contain the end time if it ended due to becoming a case or censored from loss to follow up or the last date of the study. Set the status variable as per the follow up time and censored records.

2. **Step 2: Create a file sequence variable:** File sequence variable is a type of variable that identifies and represents the records in a file in an ordinal sequence. The purpose of file sequence denoted as \`fileseq\` variable is to preserve the record sequence in the dataset. When writing a syntax, such \`fileseq\` number could be used for sorting the data in the preferred sequence, creating blocks of data using indexing and slicing methods. For example, to extract every 10th record from the dataset. Here the \`fileseq\` variable plays an important role as it is used to derive sample frame variable.

3. **Step 3: Create a sample frame variable:** A sampling frame is a frame created using a set of lists of characteristics like age, gender etc; such that the frame is used for sampling from the datasets to create stratified subsets. A sample frame variable denoted as \`sampfram\` is created using the file sequence variable. A sampling frame defines the frame of individuals those are contributing to the person-time on the case's incident date and guides in control selection. The value of the sample frame variable is the file sequence variable value of the last preceding record to have a follow-up time at least one day after the current record.

4. **Step 4: Create risk set by selection from sample frame:** Assume that for a dataset, sampling frame variable is successfully created. So, for example a case with fileseq = 200 might have a sampfram = 198. This indicates that the first 198 records are available for sampling of controls. Record number 199 is not available because it exited the person-time at risk on the same follow-up day as the case. If five controls need to be selected, five uniform 0-1 random numbers are generated and each is multiplied by 199 with the results rounded using the flooring function to the nearest whole numbers. If the same number is drawn twice, a redraw takes place. These five numbers are then the fileseq values of the randomly selected controls in the risk set for that case.

Advantages:

1. For example, in time restriction strategies operation on outcome ascertainment, it is possible to use nested case control study with incidence density sampling. Incidence density sampling proves to be one of the remedies in survivor treatment bias. Survivor treatment bias also known as time-dependent bias or immortal time bias is a type of selection bias that has the nature such that, the patients in the cohort who survive longer have more opportunity to be exposed to treatment. Nested case control study with incidence density sampling could reduce this type of selection bias. Other types of sampling such as random sampling and match sampling, will lead to bias.
2. Another example, to calculate and estimate the ratio of exposed to non-exposed controls including the person time, incidence density sampling is used. The relationship is used to estimate the incidence rate ratio (IRR) that compares the rate of exposed with the rate of non-exposed. Thus, mathematically, it is expressed as

$$IRR = \ ad/bc$$

$$where,$$

$$ad/bc - cross\ product\ or\ the\ odds\ ratio(OR)$$

While incidence density sampling could be used for calculating the IRR, other type of sampling does not support the inclusion of person time and partitioning of the risk sets that provides more nuanced results on the effect measures.

4. You and your colleagues are planning a study to investigate the risk of cardiac death in patients who have discontinued their statin medications (to treat hypercholesterolemia) after 12 months of treatment compared with patients who persisted with statin therapy for 5 years. You have an open/dynamic population of statin users with data from 01/01/2000 to 31/12/2019.

   You are concerned that the study design could introduce a particular type of bias. Outline what type of bias might be problematic for this study and describe the process you would recommend to your colleagues to remove this bias as an issue for the study. **(5 marks)**

   Solution:
   There are different types of bias that could occur for this study. Some of them are:

   1. Design bias: The selection of study design either observational study design or nested case control cohort study design depends on the experience and preference of the researcher. Because of this, there is a possibility that design bias could occur. A design bias occurs when the research study design is influenced by the preferences of the researcher.
      Recommendation:
      It is recommended that to avoid the design bias, it is always better to consult with experts in the study design for the appropriate suitability with the research study. A form of pilot testing with the selected study design should be conducted before an actual and full-fledged design implementation. This would help in confirming whether the selected study is a right fit for the research study and potential could save time, efforts, and cost.

   2. Selection bias: The type of bias causes a distortion in the estimate of the effect measure resulting from the process or way the subjects are selected from the study population. Furthermore, there are three subtypes of selection bias that are relevant to the specific study design. They are,
      a. Detection bias: This type of selection bias is induced by Sampling frame and ascertainment biases. In this type of bias, the cases of diseases are more likely to be detected in persons who tract the disease due to anticipation of the said disease.
         Recommendation:
         To adjust and mitigate the bias, the remedy is to restrict the domain of the study to a primary study base with more homogenous case detection efforts.
      b. Survivor treatment bias: This type of selection bias is induced by sampling frame and ascertainment biases. Survivor treatment bias also known as time-dependent bias or immortal time bias; is identified when the patients who survive longer have more opportunity to be exposed to treatment.
         Recommendation:
         To resolve this type of bias in data linkage studies, considering the study to be a nested case control study, it would be ideal to use incidence-density sampling for more stratified selection of case-controls from the cohort.
      c. Loss from follow up: This type of selection bias arises during study implementation. The bias is caused by exclusion of patients due to loss from follow-up or withdrawal related to either exposure or outcome.
         Recommendation:
         To mitigate to this type of bias, it is suggested to improve data linkage tracing systems. Restrict study population to a low migration group. Another option is to perform sensitivity analysis to assess the impact of different assumptions about the selected participants on study results.

   3. Information bias: The type of bias that causes flaws in measuring exposures or outcomes. There are sub types of information bias that are relevant to the specific study design. They are.
      a. Differential misclassification bias: The type of bias is caused due to misclassification of diseased / non—diseased difference between groups. This type of bias is a common and ubiquitous problem in the data linkage studies. The bias could occur in either direction.
         Recommendation:

The remedy to this type of bias is to choose and use more accurate measurements.

b. Non-differential misclassification bias: The type of bias is caused due to misclassification of diseased / non – diseased similarity across the groups. This type of bias is common in data linkage studies. The bias could be towards the null for binary data.
   Recommendation:
   The bias could be resolved by using more accurate measurements.
c. Recall bias: This type of bias is a form of misclassification bias. Recall bias is further categorized into three types. First, rumination bias which is identified as the presence of the outcome influencing the perception of its cause. Second, exposure suspicion bias identified when the presence of the outcome influences search for diseased state. Third, participant expectation bias identified when the expected diseased state may influence perception of the outcome.
   Recommendation:
   To resolve these types of biases, blind the participant as to the study hypothesis. Such type of bias is unusual in data linkage studies.
d. Hawthorne effect: This type of information bias is identified when there is a change in outcome that is caused by knowledge of being observed. This bias is unusual in data linkage studies.
   Recommendation:
   There is no specific remedy to resolve this, but only suggestion is that caution should be observed in publicising results.
e. Lead time bias: This type of information bias is identified when follow-up of two groups does not commence at comparable times.
   Recommendation:
   To avoid such biases, the need to ensure equivalent time-zero for follow-up should be enforced.

5. The Department of Health has commissioned you to undertake an evaluation of the annual prevalence of Chronic Obstructive Pulmonary Disorder (COPD) from 2010-2019 in Western Australia. You have been given a platform file comprising whole-population administrative hospital separation records (*HMDSdata*), dispensed medicine claims (*PBSdata*), and death registrations (*Dthdata*).

   Assume that your platform file is a type I dataset and has the following variables and covers the period 1st January 2005 to 31st December 2019:

Rootlpno:	unique person ID
Disp_date:	date of first known dispensing for a COPD-related medicine (blank if none)
Diag_date:	date of first hospital diagnosis record indicating COPD (otherwise blank)
Death:	date of death (blank if still alive at the end of the observation period)

   Write out documented syntax (in SPSS, SAS, Stata or R) that you would use to perform the following tasks:

   i. Determine the total number of prevalent COPD cases from 01/01/2010 - 31/12/2019.

   ii. Determine the number of prevalent COPD cases at 30/06/2010.

   iii. Calculate the change in-point prevalence from 30/06/2010 to 30/06/2019. **(15 marks)**

Solution:

1. Preprocessing
   The first step is to perform preprocessing on the provided data file. Following are the steps that would be considered for preprocessing.
   a. Create the COPD condition variable given the ICD - 9 codes:
      Assuming that in the given data file, the diagnosis code columns are available, and the diagnosis codes follow ICD-9 coding standards, first create a vector with the relevant clinical coding. Second, create an indicator variable “copd”, initialized to a value of 0. Third, using the vector filter set the “copd” indicator variable.

```
# Create a vector list that contains the desired ICD – 9 codes
codes <- c(491, 492, 493, 494)


# Create an indicator variable
data <- data %>% mutate(copd_ind=0)

# In the data frame for the given diagnosis code columns, loop through the diagnosis
# code columns and check if the given diagnosis code is present in the column. If
# present then set the indicator variable as 1 else 0
for(i in 1:21){
    data <- data %>%
        mutate(copd_ind=ifelse(!!sym(paste('diag', i, sep='')) %in% codes,
                               1, copd_ind))
}

# the !!sym is an implementation of diffusing expression in rlang along with symbol
# function that dynamically lets the mutate function access diagnosis columns created
# as string to the data frame
```

   b. Calculate the exit time based on the death:
      Once the COPD patients are identified, it is important to assign the exit date based on the death of the patient. By assigning exit date, it makes it possible to censor the patient that are relevant to the study.

```
# using the file above
# as the file already consists of death information
data <- data %>%
       # create the exit variable
       rename(exit=death) %>%
       # create the dead indicator variable using the exit column
       mutate(dead=ifelse(!is.na(exit), 1, 0)) %>%
       # change the exit column using the dead and the study end date reference
       mutate(exit=as.Date(ifelse(dead == 0 | exit >= as.Date(“2019-12-31”),
                                   “2019-12-31”, as.character(exit)))
```

c. <u>Identify COPD stages:</u>
As per internet research there are 4 stages in the COPD as a disease. It is useful to consider the disease stages for COPD that would provide a more granular level of information. The details about stages could be used to calculate prevalence of the disease at every stage. Assume the stages column for COPD are previously created. Now to create the stage person time partitions consider the below code

```
# using the above file
data <- data %>%
       # for stage 4-person time partition
       mutate(stgpt4=ifelse(!is.na(stage4), exit-stage4 + 1, 0)) %>%
       # for stage 3-person time partition
       mutate(stgpt3=ifelse(!is.na(stage3), exit-stage3 + 1 – stgpt4, 0)) %>%
       # for stage 2-person time partition
       mutate(stgpt2=ifelse(!is.na(stage2), exit-stage2 + 1 – stgpt3, 0)) %>%
       # Calculate the total partition time
       mutate(totalpt=as.numeric(exit-date + 1)) %>%
       # for stage 1 person time partition
       mutate(stgpt1=totalpt – stgpt4 – stgpt3 – stgpt2) %>%
       # Sort the data as per the rootlpno and date
       arrange(rootlpno, date)
```

d. <u>Identify if there are negative person time and fix them:</u>
It is possible that due to data recording or systematic errors to have incorrect date captured and the partition for the person time at every stage could be negative. It is important to identify the negative person time partitions and rectify it.

```
# identify the negative stage person time
identify_negative <- function(data_frame, name){
       #’ @description a function to identify the negative values
       #’ @param data_frame: data frame to use
       #’ @param name: name of the column to filter and check
       # filter the given column with negative values and get the count
       count_val <- data_frame %>%
              filter(!!name < 0 ) %>%
              count()
       # access and assign the count value
       count_val <- count_val[[1]]
       # display the count corresponding to the column
       sprintf(“there is / are %s negative record(s) in %s ”, count_val, name )
}
```

```
# Identify negative values in respective columns
identify_negative(data, expr(stgpt1))
identify_negative(data, expr(stgpt2))
identify_negative(data, expr(stgpt3))
identify_negative(data, expr(stgpt4))
identify_negative(data, expr(total_pt))

# to fix the negative person time first modify dead and exit columns
data <- data %>%
       # modify the dead indicator variable for negative person time values
       mutate(dead=ifelse(total_pt < 0 | stgpt1 < 0 | stgpt2 < 0 | stgpt3 < 0 |
stgpt4 < 0, 0, dead)) %>%
       # modify the exit variable as per the negative person time values
       mutate(total_pt < 0 | stgpt1 < 0 | stgpt2 < 0 | stgpt3 < 0 | stgpt4 < 0,
              as.Date(“2019-12-31”, origin=”1970-01-01”), exit, origin=”1970-01-
01”))


# recalculating the person time partitions
data <- data %>%
       # for stage 4-person time partition
       mutate(stgpt4=ifelse(!is.na(stage4), exit-stage4 + 1, 0)) %>%
       # for stage 3-person time partition
       mutate(stgpt3=ifelse(!is.na(stage3), exit-stage3 + 1 – stgpt4, 0)) %>%
       # for stage 2-person time partition
       mutate(stgpt2=ifelse(!is.na(stage2), exit-stage2 + 1 – stgpt3, 0)) %>%
       # Calculate the total partition time
       mutate(totalpt=as.numeric(exit-date + 1)) %>%
       # for stage 1 person time partition
       mutate(stgpt1=totalpt – stgpt4 – stgpt3 – stgpt2) %>%
       # Sort the data as per the rootlpno and date
       arrange(rootlpno, date)
```

2. <u>Prevalent cases of COPD cases from 01/01/2010 - 31/12/2019:</u>
    a. Filter the data as per the given dataset and create the prevalence column as per the dates and the stages

```
start_date <- “2010-01-01”
end_date <- “2019-12-31”

subset_data <- data %>%
       # assign prevalence indicator given the dates
       mutate(prevalence=ifelse(data <= start_date & exit >= end_date, 1, 0)) %>%
       # assign prevalence as per stage 4
       mutate(prevalence=ifelse(!is.na(stage4) & stage4<=start_date & exit>=end_date,
4, prevalence)) %>%
       # assign prevalence as per stage 3
       mutate(prevalence=ifelse(!is.na(stage3) & stage3<=start_date & exit>=end_date,
3, prevalence)) %>%
       # assign prevalence as per stage 2
       mutate(prevalence=ifelse(!is.na(stage2) & stage2<=start_date & exit>=end_date,
2, prevalence)
```

    b. Display the prevalence at each stage level

```
table(subset_data$prevalence)
```

3. <u>Prevalence COPD cases at 30/06/2010:</u>
    a. Filter the data as per the given dataset and create the prevalence column as per the dates and the stages

```
start_date <- “2010-06-30”
end_date <- “2010-06-30”

subset_data <- data %>%
       # assign prevalence indicator given the dates
       mutate(prevalence=ifelse(data <= start_date & exit >= end_date, 1, 0)) %>%
       # assign prevalence as per stage 4
       mutate(prevalence=ifelse(!is.na(stage4) & stage4<=start_date & exit>=end_date,
4, prevalence)) %>%
       # assign prevalence as per stage 3
       mutate(prevalence=ifelse(!is.na(stage3) & stage3<=start_date & exit>=end_date,
3, prevalence)) %>%
       # assign prevalence as per stage 2
       mutate(prevalence=ifelse(!is.na(stage2) & stage2<=start_date & exit>=end_date,
2, prevalence)
```

b. Display the prevalence at each stage level

```
table(subset_data$prevalence)
```

4. Change in point prevalence from 30/06/2010 to 30/06/2019
   a. Filter the data as per the given dataset and create the prevalence column as per the dates and the stages

```
start_date <- “2019-06-30”
end_date <- “2019-06-30”

subset_data2 <- data %>%
       # assign prevalence indicator given the dates
       mutate(prevalence=ifelse(data <= start_date & exit >= end_date, 1, 0)) %>%
       # assign prevalence as per stage 4
       mutate(prevalence=ifelse(!is.na(stage4) & stage4<=start_date & exit>=end_date, 4,
prevalence)) %>%
       # assign prevalence as per stage 3
       mutate(prevalence=ifelse(!is.na(stage3) & stage3<=start_date & exit>=end_date, 3,
prevalence)) %>%
       # assign prevalence as per stage 2
       mutate(prevalence=ifelse(!is.na(stage2) & stage2<=start_date & exit>=end_date, 2,
prevalence)
```

   b. Display the change in prevalence at each stage level

```
# This would give the difference between the prevalence at dates 30/06/2019 and
# 30/06/2010
table(subset_data2$prevalence) – table(subset_data1$prevlance)
```

6. You have been asked by the Cancer Council to undertake an evaluation of existing data to determine if men with prostate cancer who receive radiotherapy have improved 5-year survival compared with those who undergo radical prostatectomy (ie, complete removal of the prostate).

   You have available 25 years of linked administrative health data for the population of Western Australia, consisting of Medicare Benefits Scheme (MBS) claims (filename: *MBSdata*), hospital separation records (*HMDSdata*) and death registrations (*Dthdata*).

   Given the nature of your available data resources, explain how you would tackle: i) the ascertainment of the ***'exposure'***; ii) the ascertainment of the ***'outcome'***; iii) an appropriate ***'domain'*** for the study sample and research questions; and iv) an appropriate form of study '***design***' to answer the research question. **(15 marks)**

Solution:

Given the data and problem statement each point could be tackled as follows:

A. Ascertainment of exposure
   a. It is possible to select the men who are diagnosed with prostate cancer in the given 25-year frame based on the ICD codes from the HMDS data file. An algorithm would be devised using the clinical codes to identify people. The decision could be endorsed by consultation with practising clinicians, pharmacists and other health professionals working in relevant cases.
   b. To select cohort of men who are diagnosed with prostate cancer and have received radiotherapy post diagnosis and have survived for 5 years the HMDSdata, MBSdata and Dthdata could be merged to an ultimate file and used for selecting the sample. A secondary analysis platform file, which is a precursor to the ultimate file could be used to undertake a capture-recapture analysis of completeness of case ascertainment. Capture-recapture methods could be used for correction of case under-ascertainment.
   c. Ascertainment fraction could be calculated where the numerator is the observed number of cases and denominator is the estimated total number of cases. The higher the ascertainment fraction the better the sample is considered. This would help in measuring the exposure.
   d. A cross tabulation of Dthdata x HMDSdata x MBSdata can be performed to obtain multi-way contingency tabular input that may be used in a log-linear analysis to estimate the number of missing cases. Chapman estimator could be used for cross checking and reinforce a positive dependence on ascertainment fraction.
   e. Time restriction strategies should be employed for exposure ascertainment. It would improve internal validity. Protopathic bias could be reduced by restricting the time interval over which health care exposure is ascertained to exclude exposure wash out just prior to time-zero to follow up. This method emphasizes the population those enrolled in the cohort at time-zero to have no history of the outcome.

B. Ascertainment of outcome
   a. As an ultimate file is created during ascertainment of exposure, the same file could be used for ascertainment of the outcome.
   b. A death conditional sequence variable could be created such that, the death sequence variable is tagged as 1 only when the cause of death is because of prostate cancer. When the cause of death is due to other comorbidities the death sequence variable is set to zero.
   c. A status variable indicating if the radiotherapy treatment was received or not, admission date, transfer sequence, separation date all the them should be considered for performing life table analysis, survival analysis, cox regression and corresponding censoring where ever required and necessary. These variables could be created using columns like admission date, separation date, death date.
   d. Time restriction strategies could be employed for outcome ascertainment. It helps in improving the internal validity. Survivor treatment bias could be reduced by restricting the

period over which the outcome is ascertained such that it is no longer correlated with the period over which exposure is ascertained.

e. Random assignment specifically related to random sampling without replacement, may be useful to reduce survivor treatment bias time restriction strategy leading to an outcome ascertainment.

f. Bias associated to information bias refers to flaws in measuring outcomes. The bias could be mitigated by understanding the type of bias that is currently prevalent on the outcome.

g. It is important to optimize the external validity or scientific generalisability with effect modification. If a restricted study domain is employed on longitudinal data the generalisation of the estimated effect to other levels of the modifier may not match especially in the cases where if there are considerations beyond the study that suggest effect modification might be evidently strong and distorting the outcome.

h. It is important to determine the confounding variables affecting the outcome and appropriate adjustments needs to be implemented to attain unbiased results.

C. <u>Study domain and Research question</u>

The domain could be defined based on the focus of the research question. The research question for the given context could be defined as survival rates between men who received radiotherapy vs men who preferred prostatectomy for the period of 5 years after the diagnosis. Following are the steps that needs to be implemented for setting up the domain:

a. Create an ultimate file from the given 3 data sources. The data sources could be linked and merged by a common almost strong identifier or using set of partial identifiers. Tag all records that mentions prostate cancer at all stages, treatment received or availed for prostate cancer, for all men falling in a desired age bucket, regional and other demographic preferences. All the records that satisfy the prior mentioned condition will be tagged as 1 else 0. Filter the dataset given the tag variable.

b. Create stage partitions using the person time to attain more granular results during the survival analysis.

c. Select the data where the records contain follow-up for at least 5 years after the diagnosis of the prostate cancer.

d. Mitigation for selection bias should be made. The mitigation method would depend on the type of bias that is induced by the chosen type of sample selection technique.

e. Perform internal validation using the domain restrictions to assure accuracy of the longitudinal study data.

f. As restrictive study domain approach is utilized remedial measures to counteract with the non-generalisable effect modification caused due to consideration beyond the study should be implemented.

g. Save the domain restricted file for reusability.

D. <u>Study design</u>

There are 2 ideal study designs for the given scenario, they are

a. Retrospective cohort study: This type of study is suitable for survival analysis between different groups. The groups could be categorized as first group where patients diagnosed with prostate cancer opted for radiotherapy and the second group where patients diagnosed with prostate cancer opted for prostatectomy.

b. Prospective cohort study: The study design is considered ideal but in comparison to the retrospective it requires more time and resources. In this study men are diagnosed with prostate cancer at the study outset, then the cohort is followed forward and treatment is tracked and survival outcomes are documented. Although this approach minimizes the bias but is resource-intensive.

Following are the reasons why other type of study are not best suited for the given scenario

a. Case-control: Case control studies are used in a scenario where the investigation factors contribute towards development of a particular disease. In the given context, the problem statement explicitly states to consider a cohort where prostate cancer is prior diagnosed. The focus then lies on the survival outcomes of radiotherapy vs prostatectomy. Due to reverse causality, i.e. the inability to determine if the factors associated with receiving a specific treatment also influence survival and due to survival bias i.e. exclusion of patients due to death before the selection of case and controls all lead to bias.

b. Nested case-control: Although nested case-control could be used for the given scenario there are certain problems associated to it. First problem is associated to selecting controls where it might be extremely difficult to identify appropriate controls within the timeframes. This could lead potentially to selection bias. Second is associated to survival bias like that of the case control.

c. Cohort crossover: This design is bested used in scenarios where a study needs to be investigated on an individual for studying the effects of different interventions on the same individual over time. It is not possible to consider comparison of patients who are in two different treatment outcomes.

d. Case-control analysis of matched pairs: By matching the cases with controls based on demographic features and other determinants like cancer stage, could prove to reduce bias in the analysis. However, the approach might be limited by the availability of suitable controls and may not capture the entire at-risk population.

End of Assignment