ASSIGNMENT 2024
TAKE-HOME QUESTIONS

Instructions:

1. *There are 9 questions with the total value of the assignment is 65 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 PMIM302 cover-sheet below before submitting.*

DEADLINE for submission of all assessment materials:

1pm on Monday 15th April 2024

Submit materials for PMIM302 - Introductory Analysis of Linked Health Data through Canvas, Swansea University

MSc/PGDip/PGCert Health Data Science
Assignment cover sheet

Swansea University
Prifysgol Abertawe
Medical School
Ysgol Feddygaeth

Module: **(Please tick the appropriate box).**	☐ PMIM102/J Scientific Computing in Healthcare ☐ PMIM202/J Health Data Modelling ☒ PMIM302 Introductory Analysis of Linked Health Data ☐ PMIM402/J Machine Learning in Healthcare ☐ PMIM502/J Health Data Visualisation ☐ PMIM602 Advanced Analysis of Linked Health Data
Module part (if applicable):	Take-home assignment
Title of assignment:	*Answers*
Student ID number:	2311233
Word count (if applicable):	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 not permitted. 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 researcher has requested an *ad-hoc* linkage of a file of death records to a file of hospital morbidity records. The researcher says that their aim is to follow the patients from the time of their first hospital admission until the correct date of death (in those who are deceased). The following are the partial identifiers in the two files.

Death records – partial identifiers

File number	Social security number	Surname	First name	Sex	Date of Birth	Postcode
1	711459	Holeman	D'Arcy	M	17.02.1955	6014
2	498230	Saville	Beth	F	11.11.1969	6034
3	958940	Holman	Gary	M	04.05.2001	6074
4	345964	Batty	Jenny	F	09.10.1980	6093
5	184444	Fosil	Emily	F	29.02.1906	6009

Hospital morbidity records – partial identifiers

File number	Social security number	Surname	First name	Sex	Date of Birth	Postcode
1	711459	Holman	Cashel	M	17.02.1955	6014
2	134444	Fossil	Emily	F	29.02.1906	6009
3	223856	Holman	D'Arcy	F	04.11.1961	6034
4	334569	Holman	James	M	01.03.1981	6069
5	334568	Holman	Emily	F	01.03.1981	6069
6	345928	Batty	Gulliver	M	01.03.1981	6069
7	223856	Holman	D'Arcy	M	04.11.1967	6123
8	945872	Batty	Ron	M	04.06.1950	6069
9	138572	Holman	Daisy	F	20.02.1920	6009
10	509485	Cooper	Clarence	M	09.03.1965	6069
11	498230	Saville	Elizabeth	F	11.11.1969	6034
12	238759	Holman	Geraldine	F	05.02.1954	6049
13	209523	Battie	John	M	04.06.1950	6069
14	209523	Batty	John	M	06.04.1975	6069
15	223856	Holeman	Ignatius	M	01.03.1961	6009
16	345964	Holman	Jenny	F	09.10.1980	6093
17	985433	Sydchrome	D'Arcy	F	05.02.1995	6038
18	295843	Holman	Geraldine	F	12.12.1935	6046
19	687345	Battie	Emily	F	17.03.1945	6029
20	333345	Holeman	Kathryn	F	20.06.1973	6069

In ≤500 words, explain how you would go about performing this *ad hoc* data linkage, using technical methods covered by this course. You should draw on examples from the data shown above, as appropriate, to illustrate your points. *(10 marks)*

Solution:

Adhoc data linkage is a form of data linkage when the data required is addressed to problem by a small group of researchers.

There are two approaches to perform the data linkage:

Approach 1: To pretend SSN (social security number) to be a unique identifier.

1. Create variables by using *generic internal loading* method, create file number represents as local id. By using *generic external loading method* create a 'deathseq' variable, where the death records table acts as a lookup table using the SSN to search the record. Assign 1 if the SSN is found else 0. Separate the identifiers:
2. By *separation principle*, local id and deathseq columns are separated from the partial identifiers. The partial identifier file contains the local id along with other columns. The separated file would contain only the local id and the deathseq variable. The separated file is sent to the researchers and

the links are temporarily stored. In some cases, the file that is sent to the researchers might also contain variables like age, sex and other relevant variables that depends on the type of analysis.

3. Anonymize the identifiers:

   If the researchers require demographic partial identifiers like the date of birth, postal code for geographic / region-based research or data research, it can be anonymized or masked using differential privacy algorithms. Then the anonymized clinical data is sent across to the researchers.

Approach 2: It is possible that a person can have more than 1 SSN (1). Thus, SSN could be considered as partial identifier or at-most a strong partial identifier.

1. Understand the files:

   There are 6 partial identifiers. It is observed that there are surnames with minor changes in the spelling but not all values are unique.

2. Create ‘deathseq’ variable:

   To create a ‘deathseq’ variable, *generic internal loading* is performed. An additional column that represents the death event of the individual, where 1 represents death event.

3. Link the files:

   From step 1 the nature and application of the data to be linked is understood and thus it can be determined that, *deterministic fuzzy matching method* would be a good fit for data linkage. Deterministic fuzzy matching compares the approximate value of each partial identifier and considers a match if all the matches are resulting to be true with a user-defined error margin.

   For example, consider the sample where file number is 1, the SSN, date of birth, postcode, sex identifiers are an exact match, the surname is an approximate match and the first name is not a match. To improve the accuracy of the match, *phonetic compression method* can be used on the surnames and names so that the matching string contains only the consonants. Thus, the linking of the files is performed.

4. Create variables - local id:

   Using *generic internal loading*, local id representing the records in the merged data is created. During merging, the ‘NaN’ values will be generated for the ‘deathseq’ variable column where records mismatched and that would be replaced by zero.

5. Perform step 2, 3 from approach 1

2. Data linkage may be performed on an *ad-hoc* basis to serve only the needs of a single research project. An example of *ad hoc* data linkage was the Australian Government's follow-up of mortality and cancer incidence in military personnel who worked in the vicinity of Australian nuclear test sites. A register of the veterans exposed to the test sites was linked as a once-off exercise to the National Death Index and National Cancer Clearing House.

   Alternatively, data linkage may be undertaken on a systematic basis with no specific research project in mind at the time that the links are created. Rather, the links are stored and later retrieved to support multiple discrete research projects as the needs arise. An example of systematic data linkage is the WA Data Linkage System.

   In ≤500 words, identify and explain the advantages and disadvantages of systematic data linkage when compared with *ad-hoc* data linkage. Write your answer in the space below, which continues to the next page. *(8 marks)*

Solution:

Advantages of SDL (Systematic Data Linkage) over ADL (Adhoc Data Linkage):

1. Broader scope of research:
   Due to the distributed and linked nature of the system it makes it possible to connect dataset from different domains. It is possible to conduct longitude research as the data is accumulated overtime and makes it possible to link old and new data and construct different data structures.
2. Commercial and competitive benefits:
   SDL attracts independent researchers, other nations, and organisations to provide grants for conducting research. This increases the revenue, sustains employment and economy of the State with SDL.
3. Advance and extend existing assets:
   SDL improves the quality of administrative datasets, reduce duplication errors by providing enhanced linkage strategies, increase accuracy of recording and eases networking with the existing datasets that lead to wider range of data association.
4. Time and cost efficiency of research:
   The links stored from previous research can be updated and re-used for new research. Setting up the infrastructure is a one-time investment. Maintenance of the infrastructure is lesser than the amount required each time for setting up everything from scratch in ADL systems.
5. Privacy and Data conservation:
   SDL has decentralised architecture. SDL system can anonymize the data irrespective of the approval of the data owner. Patient records are hidden and only information relevant to the research is shared with the researchers. Common problems such as loss of data follow up, recall bias, selection bias, reporting bias are reduced and mitigated to some extend in SDL.
6. Community development:
   SDL proved to have a positive effect on the community. SDL has provided linkages over multiple domains not limiting to healthcare that help researchers look at problems with a wider perspective. This has led to major contribution to the scientific knowledge in general. SDL research projects has resulted in contribution to global healthcare and scientific knowledge via peer-reviewed journals. This has contributed in public health policy reformations and human welfare development worldwide.

Disadvantages of SDL over ADL:

1. Data size:
   Research and analysis of population and cohorts are impactful when the data size is large. If the data size is small and the research group is small ADL is preferred.
2. Availability of data resource:
   Datasets representing health events and other domains of an individual should be available to analyse the interaction, affects, association and correlation of factors affecting the individual health outcome. The dataset belonging to a smaller sample size or a subgroup of the entire population will not lead to meaningful data linkage. In such cases ADL becomes more economical to consider.

3. Continuous funding:
   In-order to design, construct and maintain efficient infrastructure that will operate the SDL would require a continuous stream of funding that is only possible for stable economic States and organizations with large seed funding. Where funding is small, ADL is considered as it the funding expenditure is optimized as per the scope of the research study.
4. Advanced Technology:
   Advanced encryption technologies, data anonymization techniques, data de-identification methodologies, derivation of complex differential privacy algorithms are required for managing and maintaining the sensitive data. Funding is required to acquire engineers with quality talent and state of the art computation facilities. ADL is a cheaper alternative.
5. Organization support:
   Support and approval of the State authority and organizations are required to envision, plan, setup SDL facilities. ADL is an alternative in such cases.

3. You are undertaking a linked data study to evaluate the rate of mortality from diabetic complication in a cohort of people with a diagnosis of diabetes mellitus after 40 years of age. The Data Linkage Branch has recently provided you with data from 1st January 1995 – 31st December 2014 that include the data sets and variables shown below.

Data set	Variables
Hospital inpatient data	Rootlpno
	Age
	Sex
	Race/ethnicity
	Residential postcode
	Date of admission
	Date of separation
	Separation type (ie, where discharged)
	Primary admitting diagnosis
	Secondary diagnoses (1-10)
	Primary procedure
	Secondary procedures (1-10)
Pharmaceutical dispensing data	Rootlpno
	Age
	Sex
	Residential postcode
	Drug item code
	Date of dispensing
Death data	Rootlpno
	Age
	Sex
	Race/ethnicity
	Date of death
	Primary cause of death

Given the nature of your available data resources above, explain what processes you might follow to check and 'clean' the data to be comfortable that the data sets provided by the Data Linkage Branch do not have any errors and provide reliable records for the proposed study. *(10 marks)*

Solution:

Data checking and cleaning could be carried out in two approaches, inter-dataset checking and intra-dataset checking.

Inter-dataset: This type of dataset check is associated with confirming accuracy and consistency of the data between the linked datasets.

Intra-dataset: This type of dataset check is associated with confirming accuracy and consistency of the received data within the data itself.

Following are the steps, that would be considered for data check and data cleaning process in each of the above-mentioned approaches.

1. Data Field Categorization:
   Data fields of the given 3 dataset needs to be understood and the categorized into ordinal, nominal and text value statistical data types. This would help provide context during checks performed on the values of the fields. Data dictionary associated with the columns would help in understanding the possible range of values that the fields can contain. Out of all the data fields strong and weak partial identifiers needs to be determined. This would help in further decision making for selection of type of data linkage technique.

2. Data Checks:
   a. Consistency Check:
      i. Units of the values in the respective columns should be checked either with a data dictionary or with the help of a subject matter expert and as per the context of the variable. For example, rootlpno should be strictly of a unique data type, date data fields should be strictly of date data type and of the same format across datasets, age field should be of continuous data type, race column should be of string values. Age value should be greater than 40. The date range value for all the date related fields should be within the range of 1st January 1995 and 31st December 2014.
      ii. The rootlpno columns looks like a unique identifier, a check for overlapping records between 3 records should be performed to ensure data for respective individuals are available.
      iii. The period for which the 3 datasets are available should be overlapping. The date of admission, separation, item dispensing, death should be sequential and in the approximate range. The death event date or separation date of any individual should not precede the admission date. In case of death event, difference between admission date and death date should not be absurdly large. The admission date and separation date should not be absurdly large.
      iv. Cohorts with respect to identifiers rootlpno, sex, age, postal code should be overlapping across the datasets.
      v. The conditional values like the primary cause of death variable should be checked for consistency. For example, primary cause of death variable should only exist in the event of death only. Similarly, date of dispensing should be present when drug item code is present and vice versa.
      vi. Diagnosis and procedure columns should contain valid diagnosis and procedure codes. It should be cross checked in reference to valid standards like ICD, CPT, SNOWMED, UMLS clinical coding.
   b. Descriptive Statistics:
      Descriptive statistics like mean, median, mode, variance, standard deviation, range, quartiles can be used to understand the distribution and nature of the data. Comparison of the data distribution would help uncover any inconsistencies in the dataset.
   c. Missing Data Analysis:
      Missing data can be analysed by understanding its type. There are three major types of missing data, missing at random (MAR) – missingness of the data associated with other columns of the data, missing completely at random (MCAR) – missingness of the data is completely random, missing not at random (MNAR) – missing of the data is associated with other variables that are not considered in the datasets. Possibility of an imputation can be performed post determining the type of missing data.
   d. Outlier Analysis:
      Check for outlier using the inter-quartile range (visuals of box plots), compare with acceptable range values with data dictionary, would help identify extreme values that could skew and add bias to the results.

4. Describe the purpose of each of the following sequence variables that can be created for use with managing linked data files. Illustrate your answer with an example of where each variable may be useful in the analysis for a specific research study. *(3 marks)*

    - Fileseq
    - Morbseq
    - Indexseq

Solution*:*

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 `fileseq` variable is to preserve the record sequence of the data. When writing a syntax, such `fileseq` number can be used for sorting the data in the preferred sequence, creating blocks of data using indexing and slicing methods. For example, to extract every 10^{th} record from the file

```
# create fileseq variable
data <- data %>% mutate(fileseq=row_number())
# select as per file sequence variable
data %>%  filter(file_seq %% 10 == 0)
```

Morbidity sequence variable:
Morbidity sequence variable is a type of block variable that identifies and represents the ordinal sequence of an individual in the file that corresponds to the morbidity of the individual. The purpose of the morbidity sequence variable is to identify the number of times an individual data was recorded due to the morbidities of that individual. Consider the following example, where the data to be selected are first-time admissions.

```
# code to select only those records with first morbseq
data %>%
   groupby(unique_id) %>%
   filter(morbseq == 1)
```

Index sequence variable:
Index sequence variable is a type of block variable that identifies and represents a sequence corresponding to a given health condition. Index sequence is sequence of integer values. The purpose of index record from the index sequence is to highlight the first record that mentions a health condition. Consider the example where index record across individuals is extracted for survival analysis.

```
# code to select only those records with index records
data %>%
   filter(indexseq == 1)
```

5. Inter-hospital transfers are often an issue for analysis of hospital morbidity data.

- In what circumstances do they need to be taken into account? *(2 marks)*

- Outline the steps using statistical syntax that can be used to address the issue of inter-hospital transfers in hospital data. *(3 marks)*

Solution*:*

Inter-hospital transfers should be considered when

1. Determining the Length of stay in hospitals:
   Length of stay determines the period when the patient was admitted in the hospital for care. If the admission dates, readmission dates and separation dates are not identified correctly, the length of stay would be incorrectly calculated. So, it should be calculated as an aggregated or cumulative measure of period of each hospital episodes.
2. Determining health outcomes:
   Procedures and conditions are significant when modelling and determining for health outcomes. So, the procedures or conditions that has occurred during the transfer period needs to be accounted.
3. Determining Risk of readmission:
   Post transfer admissions should not be counted as readmissions. So, appropriate adjustments need to be made to time-zero variable from the last separation date of a transfer cluster when determining the risk of readmission.

Following is the syntax that can be used to address the issue of inter-hospital transfers in the hospital data,

1. Create a morbseq variable to determine the file time admissions, where first time admissions would morbseq = 1.

```
data <- data %>%
  group_by(rootlpno) %>%
  arrange(sepdate) %>%
  mutate(morbseq=row_number()) %>%
  umgroup()
```

2. Create transseq variable to identify inter-hospital transfers sep_type =2 and admdate < LAG(sepdate).

```
data <- data %>%
  arrange(rootlpno, sepdate) %>%
  group_by(rootlpno) %>%
  mutate(transseq=ifelse(
    (morbseq >= 2 & lag(sepdate) > admdate) |
      (morbseq >= 2 & lag(sepdate) == 2 & lag(sepdate) == admdate), 1 0)) %>%
  ungroup()
```

3. Adjust the transseq variable to account for morbidities of the patient

```
data <- data %>%
  group_by(rootlpno) %>%
  mutate(transseq=ifelse((morbseq >= 2 & transseq == 1 & !is.na(lag(transseq)) &
lag(transseq) >= 1),lag(transseq) + 1, transseq)) %>%
  mutate(transseq=ifelse((morbseq >= 2 & transseq == 2 & !is.na(lag(transseq)) &
lag(transseq) >= 1), lag(transseq) + 1, transseq)) %>%
  mutate(transseq=ifelse((morbseq >= 2 & transseq == 3 & !is.na(lag(transseq)) &
lag(transseq) >= 1),
    lag(transseq) + 1, transseq)) %>%ungroup()
```

4. Implement backflow separation to create final separation date for the records that are part of the transfer set.

```
data <- data %>%
  group_by(rootlpno) %>% arrange(sepdate) %>%
  mutate(findate=NA) %>%
  mutate(findate=as.Date(
    ifelse(transseq == 4 & row_number() == n(), sepdate,
           ifelse(transseq == 4 & n() > 5 & lead(transseq) == 5,
                  lead(findate),
                  ifelse(transseq == 4, sepdate, findate))),
    origin='1970-01-01')) %>%
  mutate(findate=as.Date(
    ifelse(transseq == 3 & row_number() == n(), sepdate,
           ifelse(transseq == 3 & n() > 4 & lead(transseq) == 4,
                  lead(findate),
                  ifelse(transseq == 3, sepdate, findate))),
    origin='1970-01-01')) %>%
  mutate(findate=as.Date(
    ifelse(transseq == 2 & row_number() == n(), sepdate,
           ifelse(transseq == 2 & n() > 3 & lead(transseq) == 3,
                  lead(findate),
                  ifelse(transseq == 2, sepdate, findate))),
    origin='1970-01-01')) %>%
  mutate(findate=as.Date(
    ifelse(transseq == 1 & row_number() == n(), sepdate,
           ifelse(transseq == 1 & n() > 2 & lead(transseq) == 2,
                  lead(findate),
                  ifelse(transseq == 1, sepdate, findate))),
    origin='1970-01-01')) %>%
  mutate(findate=as.Date(
    ifelse(transseq == 0 & row_number() == n(), sepdate,
           ifelse((transseq == 0 & n() > 1 & lead(transseq) == 1),
                  lead(findate),
                  ifelse(transseq == 0, sepdate, findate))),
    origin='1970-01-01')) %>%
  ungroup() %>%
  arrange(rootlpno, sepdate)
```

5. After step 4, it becomes possible to calculate LOS as follows

```
data <- data %>%
   # Calculate LOS.
   mutate(los=sepdate-admdate) %>%
   # Account for LOS=0 -> 1.
   mutate(los=ifelse(los==0, 1, los)) %>%
   # Calculate total LOS.
   mutate(totlos=findate-admdate) %>%
    # Again, set zero LOS to 1.
   mutate(totlos=ifelse(totlos==0, 1, totlos))
```

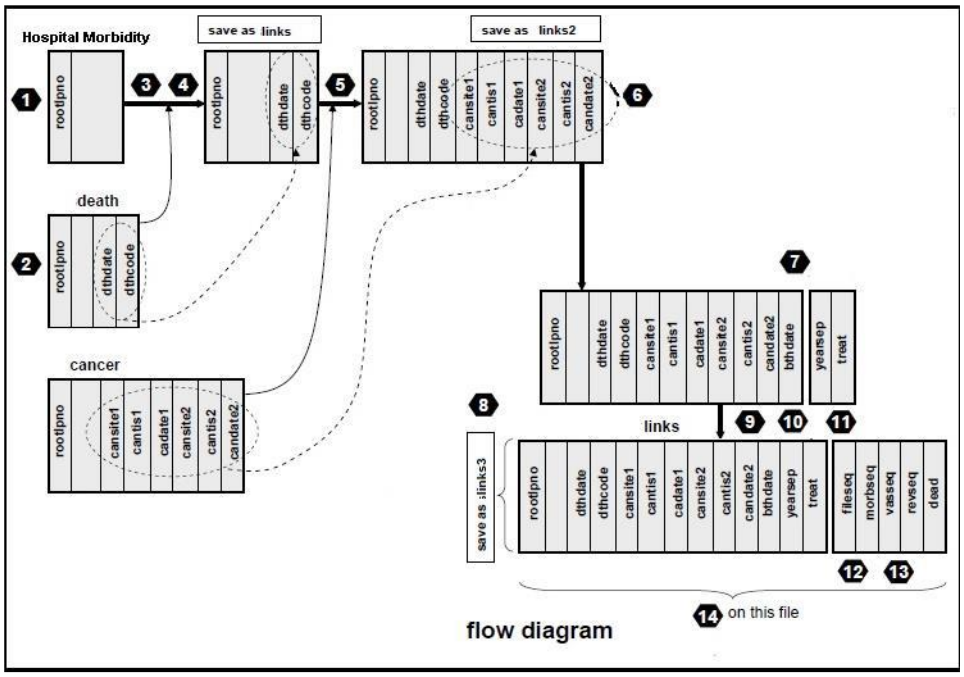
6. To ensure that the procedures code is available in the same context, need to merge two files. Assume transfer_data is a dataset that contains the procedure codes from transferred hospital.

```
data <- data %>%
  full_join(transfer_data, by=c("rootlpno" = "rootlpno"))
```

6. You are undertaking a linked data study to evaluate the rate of hospitalisation in the three years following diagnosis for a cohort of women with breast cancer who were diagnosed after 40 years of age. The Data Linkage Branch has recently provided you with a full abstract of data from three linked data collections: i) Hospital Morbidity Data System, ii) Cancer Registry, and iii) Mortality Register for the period 1st January 1995 – 31st December 2010.

   Please draw an annotated diagram to illustrate how you would go about managing and 'manipulating' your three original data files so as to produce an appropriate file for analysis of the research question above. Your diagram should contain sufficient detail to demonstrate how any domain restrictions would be applied, how files will be merged, what variables would be required from the original data sets, what new variables would need to be created in the end-of-record loading area, as well as how relevant outcome information would be ascertained.

   (9 marks)

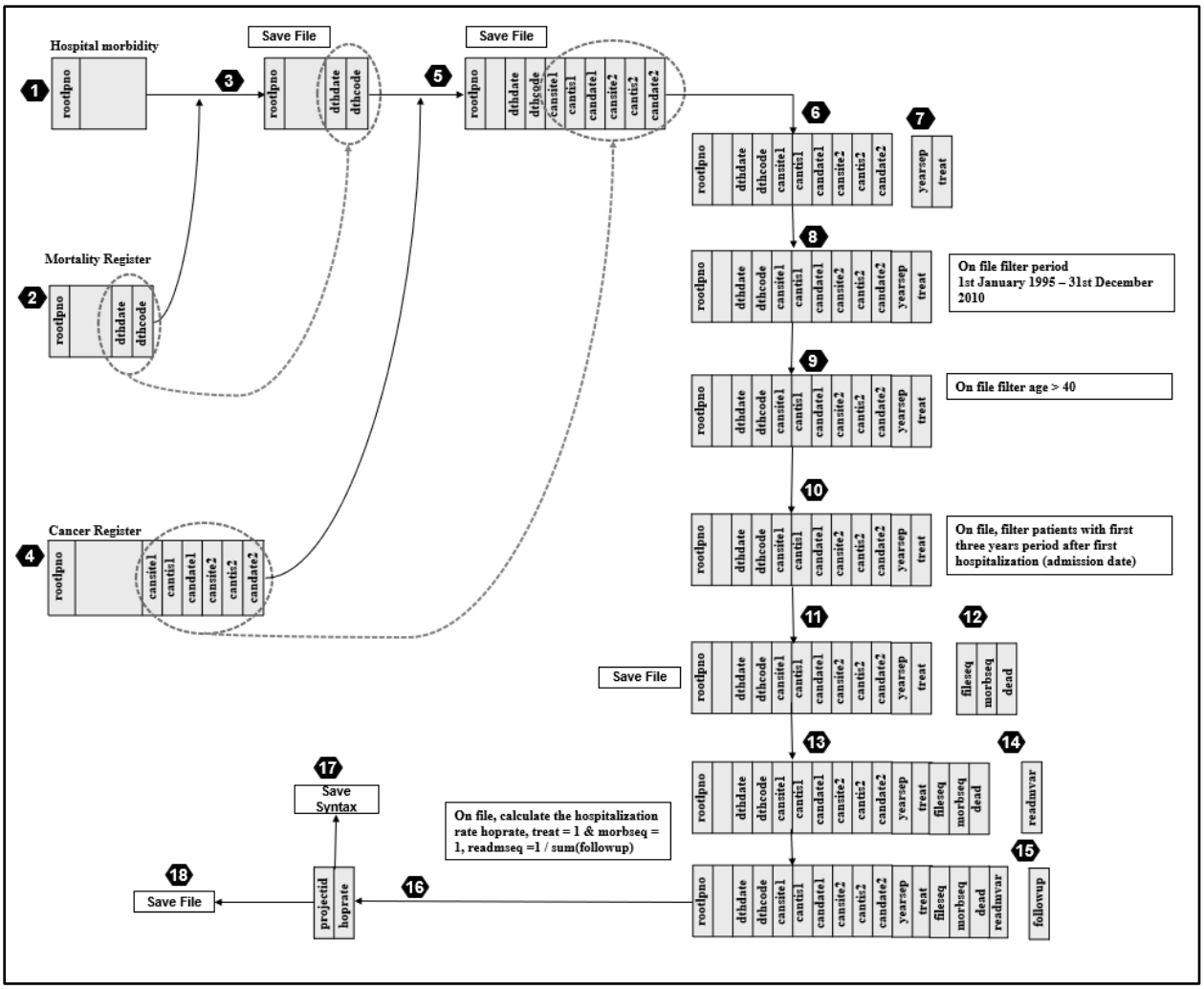


Solution*:*

Following are the steps to calculate rate of hospitalization:

1. Load the *hospital morbidity* file
2. Load the *mortality register* file
3. Merge the *mortality register* file and *hospital morbidity* file on the unique identifier '*rootlpno*' such that new columns '*dthdate*' (death date) and '*dthcode*' (death code) are added to the morbidity file. Save the merged file. Name it as '*hospital_morb_dth*' file.
4. Load the *cancer register* file.
5. Merge the *cancer register* file with *hospital_morb_dth* file such that the columns related to cancer data are available in one file. Save the merged file. Name it as '*hospital_morb_dth_cancer*'
6. Load the *hospital_morb_dth_cancer* file.
7. Extract year of separation variable named as '*yearsep*' from the separation date field and create treatment variable named as '*treat*' where the variable is set to 1 if the patient is subjected to treatment else 0.
8. On the current file, filter the data to ensure the '*admission date'* and '*separation date'* falls between the 1st January 1995 and 31st December 2010. Check for discrepancies such as the *separation date < admission date, separation date – admission date = absurdly large period.*
9. On the current file, check for distribution of '*age'* given the condition of treatment. Resolve any discrepancies. Discrepancies can include missing data, outlier data. Filter the data to ensure the '*age'* of the patient is greater than 40.

10. On the current file, filter the data for each patient to subset only the first three years of record from the '*admission date'*.
11. Save the file. Name it as '*processed_data*'.
12. Create '*fileseq*' where the values represent the sequential order of the records. Create '*morbseq*'- morbidity sequence where the value represents the sequential order of the morbidities records in a block belonging to single patient.
13. Create '*readmvar*' – readmission variable, where the value is set to 1 whenever a readmission has encountered. Create 'dead' variable where the value is set to 1 in the patient death event. Save the file. Save the file. Name it as '*processed_data*'.
14. Create a '*followup*' variable that tracks the number of days between the last separation date and readmission date.
15. On the current file, calculate the '*hoprate*' – rate of hospitalization where the numerator is the count of records given the condition *treat = 1 & morbseq = 1 & readmvar = 1* and denominator is the total person days of '*followup*'.
16. Create '*projectid*' and '*hoprate*' for reporting.
17. Save the data files.
18. Save the syntax.



7. The separation principle is often considered best-practice in data linkage. Describe this process and the advantages it provides. *(4 marks)*

Solution*:*

Process of separation principle are as follows:

1. Field identification
   After the identity files are received, identifier fields are understood. Data standardisation methods are performed.
2. File linkage
   Based on the data identifiers data linkage method is determined and performed.
3. Add project id
   A project id field is created to represent the research study for which the data is prepared.
4. Create project id local id files
   Local file id is created and linked with the project id.
5. Strip the identifying data
   Identifiers are stripped away after the linking is complete. The removal of identifiers depends upon type of analysis. Two files are created. The project file, with project id and clinical data and the identifier data file. The identifier data file is then stored locally and the clinical data file is shared with the researcher.
6. Save the links and share the clinical data
   The project id – local id link file is saved and the project file containing the clinical data is forwarded to the researcher group or organization.

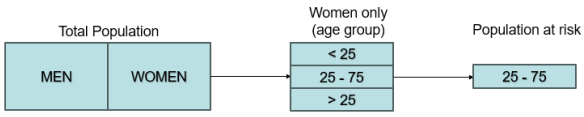
Advantages of separation principle are as follows:

1. Data conservation
   The privacy of the individuals is protected. The directly identifiable data, not so overtly identifiable data is separated from the content data. The researchers receive only an id associated with the content data.
2. Ease of data access
   Due to the separation principle, it became possible to access data without need to worry about breach of privacy. The data sharing has become possible even for those records where the data consent is not provided as the private data is not shared.
3. Time and cost efficient
   As data is shared without the private identifiers, the time and manual efforts required to get data consent is saved. It became possible to a perform research study on a wider scope of study without worrying about losing the privacy.
4. Progress in multiple domains
   Faster access to data became possible leading to increase in number of research study and overall growth and contribution towards community, wealth, and public health.

8. In a cohort study all individuals who enter the study population need to be at risk of the outcome. From a practical perspective using linked data sets, what are some things that should be implemented in the data to ensure all individuals in the final data set for analysis are part of the true 'at risk' population? *(4 marks)*

Solution*:*

The group of people who are susceptible to a given disease monitored for a period are known as population at risk. Consider the following example of study of carcinoma of cervix. Out of the total population, women of the age group 25-75 are considered as population at risk for carcinoma of cervix.



Following are some practical measures that can be considered to ensure selection of true population at risk

1. Identify the type of population
    a. Closed: When a population is selected to be monitored for a given period and the population does not change i.e. new entries of patients are not considered, then the population is a closed type of population.
    b. Dynamic: When a population is selected to be monitored for a given period changes over time i.e. new entries and exits of patients are considered, then the population is a dynamic type of population.

    It is important to identify the type of population to device suitable strategies for data management and linkage.
2. Selection of determinants
    Determinants are responsible for selection of a cohort and derivation of health outcome of interest. The wider the array of determinants the better the quality of selection of cohort and derivation of health outcome.
3. Validation of data sources
    During linkage process multiple data sources are considered. Validation of such multiple data sources through clinical guidance for data inclusion is critical for quality assurance and thereby directly improving quality of data linkage.
4. Data checks before linkage
    Need to perform data checks before data linkage. Some of the checks are:
    a. Death event: A patient to be considered in the cohort study should be alive. The records that have a death outcome event should be excluded from the study.
    b. Cannot already have the disease: Consider the given example of carcinoma of cervix, here a patient to be considered for the cohort study should not already have the disease. The population at risk expects the population to who have the chance or the probability of reaching a health outcome status.
    c. Demographic information: A sense of person, place and time and underlying health issues should align with the health outcome interest. Selection of patients in the cohort study based on the demographic information is crucial for selecting true population at risk.
    d. Consistency check: Check to confirm if the selected cohort fall in the desired period. The time of events follow a consistent sequential order. Check for missing data, outlier data and other analysis.
5. Identify Type of merging data
    There are different types of links to be selected for data linkage. One to many, one to one and many to many are examples of types of links. The types of links correspond to the merge to be selected one of either left, right or full joins / merge. This is crucial for linking and creating Type 1 and Type 2 files desirable for cohort study. Incorrect type of merging could lead to inconsistent links and records and thereby unexpected analysis outcome.
6. Data checks after linkage: Need to perform data checks after linkage. Checks include analysis of missing data, outlier data, duplicate data, merge induced errors and inconsistencies, compare the statistical distribution before and after merging.

9. You have been given a linked data file that is a merge between births (stillbirths + livebirths) occurring in 2012 and subsequent deaths, provided that the deaths also occurred in 2012. In other words, the length of potential follow-up varies from a maximum of 364 days for those born on 1 January 2012 down to zero days for those born on 31 December 2012. The linked file contains one record per birth with any deaths appended in the end-of-record loading area. The file was not in any particular order when given to you. The file contained the following variables:

rootlpno	unique personal ID.
confineno	unique confinement ID
birthdate	date of birth
status	1=livebirth; 2=stillbirth
deathdate	date of death (where applicable)

The file contains a number of multiple births, which are denoted by having the same confinement number within the same rootlpno.

You have been asked to calculate the cumulative incidence ratio, comparing members of multiple birth sets with singleton births. Write out the syntax (in SPSS, SAS, Stata or R) that you would use to generate the figures from the linked file needed to perform this calculation. Your syntax should include sufficient documentation to enable someone else to follow your reasoning and approach. *(12 marks)*

Solution*:*

Cumulative incidence (incidence proportion) for a given cohort study can be defined as the proportion of members at risk who experience a given event at the given period.

$$CI_t = a/N$$

where a is the number of people that get the condition / event in the given period and N is the population at risk. The cumulative incidence of death is called as cumulative mortality. Here, the expectation for the given data is interpreted as to find the cumulative mortality. Cumulative Incidence is of two types – Conditional Cumulative Incidence and Unconditional Cumulative Incidence. The CI of neonatal mortality is the number of deaths divided by the number of births over a period in days after birth. Following are the steps for calculating the CI ratio.

1. Load the dataset

```
#load the library
library(tidyverse)
#load the linked datafile, assuming it's in csv file format
data <- read_csv(<linked_file.csv>)
```

2. Data checks
    a. Data checks must be performed for quality assurance and consistency check.

```
# Convert dates are into the same format
data <- data %>%
   mutate(birthdate = ifelse(is.na(birthdate),  birthdate,
                       as.Date(birthdate, format = "%d/%m/%Y"))) %>%
      mutate(deathdate = ifelse(is.na(deathdate), birthdate,
                                          as.Date(deathdate, format = "%d/%m/%Y")))
```

b. Inspect and select if the dates are in the said range

```
# filter data between birthdate and death date 1st Jan 2012 – 31st dec 2012
data <- data %>%
    filter((birthdate) >= '01/01/2012' & (deathdate) <= '31/01/2012')
```

c. Inspect and select consistent and sequential birthdate and *deathdate*. This is to ensure that the consistency is maintained.

```
# filter and select only records where deathdate are greater than birthdate.
data <- data %>%
         filter(deathdate > birthdate)
```

d. Inspect and check if the status variable is correctly populated for death date present records only

```
data <- data %>%
# if death date is na and birthdate is given then set status variable to 1 as the baby
is alive
    mutate(status=ifelse(is.na(deathdate) & !(is.na(birthdate), 1, status) %>%
# if death date is not na and birthdate is given then set status variable to 2 as the
baby is stillbirth
    mutate(status=ifelse(! is.na(deathdate) &!is.na(birthdate), 2,  status))
```

3. As there are multiple birth records that have same rootlpno and different *confineno*. Such records need to be distinguished. It can be distinguished by creating the *birthtype* variable, where the *birthtype* is set to "singleton" when there are only 1 birth and set to "multiple" when there are twins or more births.

```
data <- data %>%
   groupby (rootlpno, confineno) %>%
   mutate(birthtype=ifelse(n() > 1, "multiple", "singleton"))
```

4. Create *followup* variable to determine the number of days between the birthdate and deathdate. Here since the end of study is considered as "31-12-2012" last date of follow up for the calculation.

```
data <- data %>%
   mutate(followup=ifelse(is.na(deathdate), as.Date("31/12/2012") – birthdate,
                       deathdate – birthdate)
```

5. Calculate the cumulative incidence for both type of births.

```
data %>%
   groupby(birthtype) %>%   # group by either singleton or multiple births
   summarise(cum_inc=sum(!is,na(deathdate)) / n() )
```

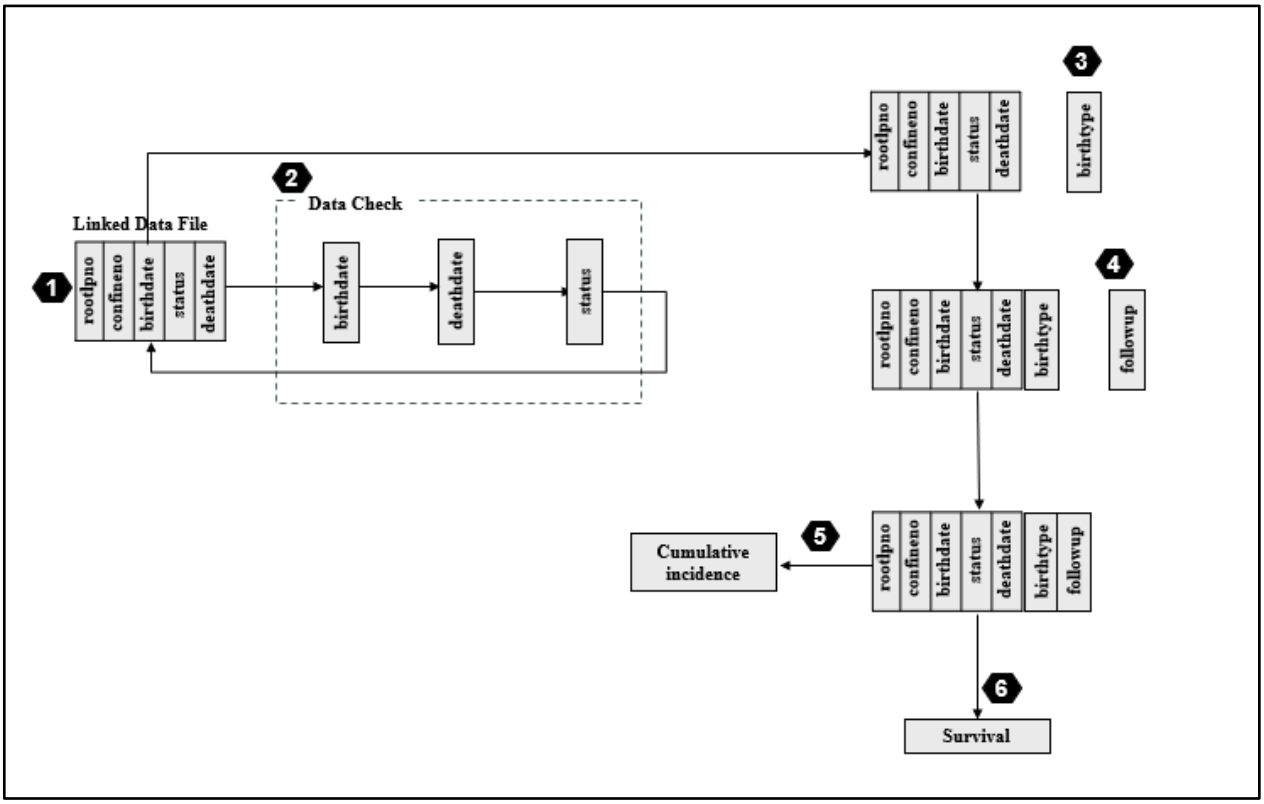
6. Calculate the survival.

```
library(survival)        # load the libraries
library(survminer)
temp <- data %>%
    select(rootlpno, followup, status)

# fit a survival object
surv <- survfit(Surv(ceiling(followup / 364), status) ~ 1, data=temp)

# plot the survival plot
ggsurvplot(surv, conf.int=TRUE, ylim=c(0.975, 1), xscale=1/364,
           title='Survival Time to Neonatal Death',
           ylab='Cumulative Survival', xlab='Days',
           legend='none')
```

Following is the annotated diagram



References:

1. SSA O. Social Security Numbers [Internet]. [cited 2024 Mar 29]. Available from: https://www.ssa.gov/OP_Home/handbook/handbook.14/handbook-1401.html

- END OF THE ASSIGNMENT-