# Introductory Advanced Linked Health Data



1. **Identify and describe each of the five stages of data linkage when using probabilistic matching.** Five Stages of data linkage

* **Preparation:** This is a vital stage of data preparation; it involves searching and correcting invalid values in data fields, frequency listing and cross-checking, recording, and transforming data into a standard format. It also produces the phonetic codes using name compression algorithms; and takes account of alternative surnames and given names for file explosion.
* **Blocking:** This step involves ordering or blocking the record files to increase the efficiency of searching for matches; it consists of sorting the files into the same order of unique identifiers. The probabilistic matching uses multiple blocking by the name of the compression algorithm and by date of birth.The blocking strategies for each pass should be independent to the extent possible. The fields consisting of the most numbers of values of possibility & highest reliability are considered and best suited for blocking*.*
* **Matching:** This process allows for potentially linkable pairs of records to be systemically compared against all other records to decide whether they relate to the same person & files of accepted links between the data sources. In Probabilistic Matching: each matching field provides some information and taken together; the fields determine the status of the pair being examined. Some fields provide reliable information compared to others, so a system of weights is used to reflect the value of comparison in each field. Each matching field contains some information and is considered; the field determines the status of the pair being examined. At the end of this matching process, a file is accepted between the data sources.
* **Storage:** Distinctive phase of data linkage system. This phase helps store the linked files from the matching phase for future use in data extraction & merging.
* **Merging:** The linkable data sets are brought together arising directly from or stored after matching phase. This phase helps in validation checks, error detection & correction of linkage errors. The data from the various sources are assembled to derive analysis as a set of composite records.

1. **In a cohort study all individuals who enter the study population need to be at risk of the outcome.**
2. **What criteria need to be satisfied for a population to be considered 'at risk'?**

Cohort population, also known as fixed population, is defined by one or more event occurrences, e.g. The population of the patients in WA first receiving the diagnosis of bladder cancer in 1980-1999. The population is 'at risk' when they are alive, secondly when they can get the disease (person, place, time), and when they don't already have the condition. Two types of risks can be considered, a cumulative incidence which is the proportion of cohort members at risk who experience a given event within a specified period of follow-up time. A rate in which the numerator is new cases of diseases or injury, and the denominator is the population at risk expressed as person-time is known as an incidence rate.

1. **From a practical perspective using linked datasets, what processes should be implemented in the data to ensure all individuals in the final data file are part of the true 'at risk' population? (2 marks)**

Cohort research is conducted over a longer period, and as a result, some subjects will be lost due to death or loss of follow-up. Cumulative incidence and incidence rate should be considered to ascertain the true risk population; apart from this, the person & time plays a vital role while identifying the population true at risk.

1. **Inter-hospital transfers are often an issue for the analysis of hospital morbidity data.**
2. **In what circumstances do they need to be taken into account? (4 marks)**

The inter-hospitality data can be of different types, i.e., serial transfers, overlapping transfers & nested transfers. In real-time, the inpatient's data are recorded as transfer records if the admission date falls on any day except the last day of another hospital or if the admission date on the separation date indicates 'inter-hospital transfer'. There are three main implications while doing an inter-hospital data analysis:

1. When determining a *length of stay* in the hospital or cumulative length of stay over a longer period, it should be considered.
2. If conditions & procedures are recorded on either the initial or subsequent records, they should be considered.
3. To avoid the readmissions as counter transfers and to adjust the time-zero for commencement of risk of readmissions from the last date of separation of a transfer cluster.
4. **Describe the steps using statistical syntax that can be used to address the issue of inter-hospital transfers in-hospital data. (6 marks)**
5. Load the hospital morbidity in R studio

Load (datafile:here:here)

1. Identify the inter-hospital transfers by Creating a ‘transseq’ variable (sepdate=2 & admdate <Lag(sepdate))

Datafile <- Datafile %>% arrange (unique-id, sepdate)

Datafile<- Datafile %>%

select (unique id, admdate, sepdate, septype, morbseq) %>%

Group by (unique id) mutate (transseq = ifelse (morbseq >2 & lag(sepdate) == admdate),1,0)) %>%

Ungroup ()

1. Setting the transseq

Set transseq =0

If morbseq >=2 & sepdate > admdate = transseq=1

If morbseq >= 2 & sepdate =2 & sepdate=admdate = transseq=1

If morbseq >=2 & transseq = 1 & transseq >=1 & transseq = transseq + 1

1. Back flow the index record & calculate the LOS
2. **Data linkage can 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 can 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.**

**Identify and explain the advantages and disadvantages of systematic data linkage when compared with *ad-hoc* data linkage.**

1. Systematic data linkage supports an indefinite number of primarily unknown objectives in future research. In contrast, the ad-hoc data linkage focuses on one research project with smaller numbers with known objectives. E.g., the Australian Government's follow up focused on morality and incidence related to cancer in military personnel only.
2. Systematic data linkage deals with unlimited datasets without any prior objectives and is more versatile and practical for, e.g., the WA data linkage system, while the ad-hoc deals with the datasets related to the objectives.
3. The data required for the systematic data linkage is partial identifiers clinical data, which is vague but can be later prioritized when objectives are framed.
4. The time of activity in systematic data linkage is a disadvantage. It requires continuous follow-ups and updates, which usually require enormous funds and human resources. At the same time, in ad-hoc, once the objectives are achieved, the time of activity can be closed, which is time effective.
5. The storage of the unit is also a disadvantage of the systematic data linkage as it requires a dedicated larger space because it stores an enormous amount of data also requires a master key linkage, while the ad-hoc data is stored as an integral part of the research and usually are just related to the study being conducted.
6. The fund's requirement for the systematic data is relatively higher when compared to the ad-hoc, which usually takes the funds directly from the distributed budget for the research and can be a disadvantage for systematic data.

1. **Index sequence variables can be a powerful tool for managing and analyzing a type 3 data file. Describe this type of sequence variable and provide two examples where an 'indexseq' variable may be useful for facilitating data management and/or analysis.**

Index sequencing variables identify the ordinal sequences of records of the same individual, but it commences with the value of 1 for the index record. The index record is the first to mention the target condition, procedure, or event that is the object of the research. If the value of Indexseq is assigned to the individual's records and is encountered prior to the index record may be assigned as blank or 0 pr sometimes; it is preferable to –1, -2, and so on, moving backwards from the record.

Example: 1- If we consider the ***HMDSdata*** set, which has data if the following variables cover the period 1st January 2005 to 31st December 2019, and if we are looking for a specific outcome, the total number of patients diagnosed with Diabetes type-II from 2015-2019. Index sequence can be added

to identify out the as 1 for the first record or first-time diabetes type II diagnosed cases.

Example: 2- In the dataset *vashmds data file* and merging the file with two others *vasdeath & vasbirth* to find out hospital morbidity, death, cancer, and birth, to work on the merged file, fileseq, morbseq should be assigned, after that as we are looking for the cases of vasovasostomy indexseq can be assigned as 1 to find out the first vasovasostomy case.

1. **Describe the three general forms of linked data research that can be performed when undertaking a longitudinal study and provide an advantage and limitation of each.**

Longitudinal data can either be an experimental study, e.g., RCT or non-experimental, e.g., cohort study. There are three general forms of a longitudinal study that can be performed using the linked data:

1. **Encapsulated Linked Data:** Identification & following up of the subjects from time zero and classification are made according to their determinant(s) or outcome(s) using the resources confined to the data sources in the data linkage system. This type of study is based on the information resources covered by the data linkage system as the only sources of data.

**Advantage:** It can help identify the changes in the determinants as it considers only a sample of the total population; additionally, it can also help measure the impacts of the change in determinants.

**Limitation:** It can be expensive, time-consuming & limited outcomes can only be determined.

1. **Linked Longitudinal Study with supplementary data:** The study participants are identified and followed up from time zero from subsequent events of interest, e.g., death or out-migration, using the database covered by the linkage system; however, this also uses the information from other databases which are not pre-linked with the data system, e.g., hospital chart reviews or doctor's questionnaires which may give some additional valuable information. The variable in the question could be a missing determinant or outcome measure, confounder, or effect modifier.

**Advantage:** The extra variable in the questions can help identify the rare outcomes, as it is done on a relatively smaller scale and could be less time-consuming.

**Limitation:** The chance of bias is high.

1. **Linked follow up of an independently enrolled cohort:** The study participants have already identified and enrolled into an experimental or non-experimental cohort population independently of the data linkage system, usually recording baseline data on confounders, exposures & effect modifiers. An ad-hoc system then is linked to a linkage system to enable follow up & ascertainment of outcomes recorded in the health databases covered by the system. This type of study would be acceptable, e.g., an RCT requiring a long term follow up for cases like stroke, cancer, or death.

**Advantages:** Multiple outcomes can be studied related to a risk factor.

**Limitation:** Expensive to conduct. Less control over the variables throughout the research.

1. **The Department of Health has commissioned you to undertake an evaluation of the hospital 'burden' resulting from type II diabetes mellitus from 2010-2019 in Wales. You have been given a merged type 3 data file comprising whole-population administrative hospital separation records (*HMDSdata*) and death registrations (*Dthdata*).**

Assume that your file has the following variables and covers the period 1st January 2005 to 31st December 2019:

**Rootlpno:** unique encrypted person ID

**Diab\_adm:** date of hospital admission indicating type II diabetes as the principal diagnosis (otherwise blank)

**Diab\_sep:** date of hospital discharge for hospital records indicating type II diabetes as the principal admitting diagnosis (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:

Load the required packages

**library(tidyverse)-** For Data Structuring.

**library(magrittr)** - For new "pipe"-like operator, %>%, with which you may pipe a value forward into an expression or function call

**library(lubridate) –** For dates and times

**library(psych)** - Multivariate analysis

**Load the HMDSdata into the R studio**

load(here::here('data/ *HMDSdata*.RData'))

head(HMDSdata)

1. **Determine the total number of patients admitted to the hospital from 2005-to 2019 with a principal diagnosis of type II diabetes.**

Number of patients with diabetes from 2005-2019:

Numdiab<-(*HMDSdata%>% group\_by (*Rootlpno*)%>%filter (! is. null (*Diab\_adm)*) %>%filter(row\_number()==1)%>% ungroup%>%summarise(totalcount=n()%>%select(totalcount)))[[1]]*

Numdiab

1. **Determine the average number of hospital admission episodes from 2005-to 2019 for type II diabetes**

Numhospadmis<-*HMDSdata%>%filter(!is.null(*Diab\_adm)*)%>% mutate(admisyear=year(*Diab\_adm*))%>%group\_by(admisyear)%>%summarise(countyear=n(), avg = mean(countyear))*

Numhospadmis

1. **Determine the total and average (per patient) length of hospital stay due to type II diabetes**

Totalaverage<- *HMDSdata%>%group\_by(*Rootlpno)*)%>%filter(!is.null(*Diab\_adm)*)*%>%mutate(daystay=difftime(Diab\_sep-Diab\_adm)%>%summarise(total=sum(daystay),average=mean(daystay))

Totalaverage

1. **Calculate the person-time at risk of death due to diabetes in your cohort.**

**Load the *Dthdata* into the R studio**

load(here::here('data/ *Dthdata*.RData'))

Data<- *HMDSdata%>%filter (!is.null (*Diab\_adm)) %>%full\_join(*Dthdata*)

Totaldeath<-(Data%>%filter(*(!is.null(*Death)) %>%group\_by(Rootlpno)*%>%filter(row\_number()==1)%>% ungroup()%>%*summarise(totalcount=n())%>%select(totalcount))[[1]]

Totalpopulation<- nrow (unique (*HMDSdata$(*Rootlpno)

Deathrisk<- Totaldeath/ Totalpopulation

***- END OF THE ASSIGNMENT-***