Statistical Analysis Plan:

Purpose/Objective:

- To study the mortality trends of liver cancer and the associated risk factors in the Laos population over a decade. (for example: 2007-2017).
- To explore the association between alcoholic cirrhosis and hepatocellular carcinoma.

Hypotheses: Non-hepatitis cirrhosis increases the risk of hepatocellular carcinoma

Authors: X, Y, Z and A

Data sources and primary variables:

Main exposure: Non-hepatitis cirrhosis risk factors (we will consider only alcohol, to keep it simple)

Main Outcome: Hepatocellular carcinoma

Other relevant variables: Other risk factors, confounders, etc.

Data sources: population registry (demographic variables), cancer registry (ICD-10 codes); medical registry (vaccinations, diagnosis, index date, treatment); prescription registry; hospital registry, death registry

Crude methods, derived variables and role of covariates

Crude analysis: would include only the main exposure

Adjusted analysis: would be adjusted for the confounders

Skeleton tables:

Baseline characteristics of the study population

Figures of the trends of mortality for liver cancer in the study population over a decade

Table giving the crude and adjusted results

Schedule: a time plan for the project (literature review; data collection and analysis; writing the article and submitting it)

Storage: storing and documenting the analysis in a directory, work server

DM:

Criteria for a good DM would include:

- An organized storage plan for the data
- The data should be well documented so it can be self-descriptive
- One should be able to track the data easily (in case with datasets)
- Even if several versions are available, a proper documentation of the versions
- Proper back-ups for all the data

For the project, a general data management plan would include,

Step 1: I would start by making a statistical analysis plan for the project. (so the aim, hypothesis, variables needed, statistical methods and software that will be used, and a project timeline). Also, at the start decide who the authors of the different scientific articles will be.

Step 2: Study the data and check for the variables that are really relevant for the project that will be used. Keep only the required data and drop the rest. Also, create self-explanatory variable names.

Step 3: Reduce space taken by the data by creating an efficient data set - avoiding redundancy of the data, reduce length of strings by labeling them, rounding up decimals, storing in UTF-8 format. Set up a denormalized data set, so it will be easy to join them when needed or only merge those needed for the analysis.

Step 4: Write codes; codes that are

Easy to understand to both self and others

Short and precise

Adding comments as to why the code is done

Creating pseudocodes

Having several versions

Step 5: Have good documentation and control of the different versions, can be done

i. manually - (directory structure, read-me files, self-explaining file name with date or version no:, templates, create back-ups)

ii. automatic - using Git for example which creates back ups on a web hotel, easy to share files with others, the files are protected, changes can be performed and all codes that are changed will be documented.