Data Processing and Analysis
4.0 DATA PROCESSING AND ANALYSIS

4.1 INTRODUCTION

Most of life and health sciences research and development functions are under increasing pressure to improve innovation, advance product safety and reduce development inefficiencies. Patient level data, collected through health record systems, offers one promising avenue for redefining Research and Development (R&D) and revolutionizing the health science value chain. Data intensive health care environment providers generate terabytes of patient data. Laboratory auto analyzers, pharmacy systems, and clinical imaging systems produce increasingly complex and voluminous data, augmented by data from systems supporting health administrative functions such as patient demographics, insurance coverage, financial data etc. Secondary uses of health data can enhance individuals health care experiences, expand knowledge about diseases and treatments, strengthen understanding of health care systems, effectiveness and efficiency, support public health and security goals and aid businesses in meeting customers’ needs.

According to “An Introduction to Secondary Data Analysis” published by Cambridge University Press, in the fields of epidemiology, public health and GIS, the distinction between primary and secondary data depends on the relationship between the person or research team who collects the data set and the person who analyzes it. This is an important concept because the same data set could be primary data in one analysis and secondary data in another. If the data set in question was collected by the researcher (or a team of which the researcher is a part) for the specific purpose or analysis under consideration, it is a primary data. If it was collected by someone else for some other purpose, it becomes a secondary data. Of course, there will always be cases in which this distinction is less clear, but it may be useful to conceptualize primary and secondary data by considering two extreme cases. In the first, which is an example of primary data, a research team conceives of and develops a research project, collects data designed to address specific questions posed by the project, and performs and publishes their own analyses of the data they have collected. In this case, the people involved in analyzing the data have some involvement in, or are at least familiar with the research design and data collection process, and know how the data were collected to answer the questions examined in the analysis. In the second case, which is an example of secondary data, the person performing the analysis did
not participate in either the research design or data collection process, and the data were not collected to answer specific research questions.

4.2 DATA INTEGRATION

GIS makes it possible to link or integrate information that is difficult to associate through any other means. It is a handy tool to perform complex analysis with spatial and non-spatial data to bring out suitable solutions for natural resource management. This mapping tool is well appreciated and incorporated in several studies considering its varied potentials in the field of resource analysis. Thus, GIS can use combinations of mapped variables to build and analyze new variables.

4.3 ENDOMETRIAL CANCER DATA AND PROCESSING

The endometrial cancer data used in this study and the related data from the Government records were used as the primary information. The incomplete records were removed and the collected data accounted to 324 endometrial cancer cases analyzed. From this data the socio-economic factors and other related information were stored in excel spread sheet which were used for further analysis.

The data with socio-economic factors were used for the logistic analysis to predict the risk factors associated with endometrial cancer other than the hereditary factors. The elaborate description of the data and methodology with results is presented in section 5.1.

The recorded cancer incidences with the population data were analyzed in the District of Mysore since the maximum cases were found in Mysore District and the Taluks also equally contributed to it. Based on this, the Google Map integrating with MapObject using Visual Basic was prepared and presented in section 5.2.

The identified risk factors in 5.1 were further analyzed using a data mining technique- Association Rule. Out of the nine parameters which were used in section 5.1, only five parameters were associated, which were further enumerated by the support and confidence values using an algorithm developed by us.

The next two sections deal with the Images collected from different hospitals. To extract the cancerous region from the MRI Images of Cancer of the Endometrium, two algorithms in Visual Basic platform were developed and presented in section 5.4. The extracted images were subjected to fractal dimension analysis using the Box
Counting Method and wavelet Analysis for classifying different stages of the cancer based on the values obtained. The results were presented in section 5.4 and 5.5.

The brief objectives and the expected outcome are presented as a flow chart in Fig 4.

Fig 4: Connections between section 5.1 to 5.5