Report: (max 4 pages w/o references)

Title: BS9001: Meta-analysis of breast cancer using data science techniques

Abstract (max 150 words) (to be done last)

**Introduction (1 to 3 paragraphs)**

Meta-analysis is essentially a statistical analysis that condenses the results of several scientific studies.1 Benefits of performing a meta-analysis is the combination of data, resulting in a higher statistical power and more robust point estimates than from individual studies.1 The steps in a meta-analysis include formulation of the research question, literature research, selection of studies, decision of which dependent variables are allowed, selection of a meta-analysis model, and examination of heterogeneity between studies.3 In this study, the stage of progress is done until the selection of studies.

Breast cancer is the development of cancerous tissue from breast tissue.4 Among breast cancers, cancers of the ducts are known as ductal carcinomas, and cancers of the lobules are known as lobular carcinomas.2 In this study, the type of breast cancer studied is known as Invasive Ductal Carcinoma or Infiltrating Ductal Carcinoma (IDC). IDC is the most common type of breast cancer, causing up to 80% of breast cancer cases.2

Cancer is a leading cause of death, with breast cancer being the most common cancer among women, totalling two million cases worldwide.2 Therefore, it is important to understand the genetic profiling of cancer among various studies to pinpoint the various significant genes that are highly associated with IDC. In this study, the aim is to obtain genetic expression data from genetic databases, to clean and manipulate it for use in the future meta-analysis.

**Methods (1 to 3 paragraphs)**

**Selection of dataset/series from the GEO database**

A search was done on the GEO online database for suitable datasets/series from studies on invasive ductal carcinoma. Any suitable datasets/series were selected, and its corresponding accession number was recorded.

**Extraction of data from database**

The gene expression profile is sourced from the GEO database via the GEO accession number and manipulated using R. The working directory is first set. The GEO GSE file is obtained and assigned to a variable. The pheno data and feature data of the eSet is assigned to their respective vectors. Expression is then obtained from the feature data via the exprs() function. The gene symbols were then sub-setted from feature data and assigned to a variable as a vector. The expression data is then combined with the gene symbol vector to form a combined data frame. Row names were replaced with gene symbols. A class factor was created for further use in the bootstrap method. A gene symbols vector was created. All data frames, factors and vectors were written into files using the write.csv() command

**Compression of Gene Expression data**

The combined data frame and gene symbols vector were first read into the R script using the read.csv() command. Using the unique() function, the gene symbol vector was removed of any duplicate genes. An empty list was created using the list() function. A simple function was written to determine the mean expression level for genes which have multiple rows. The lapply() function was used on the list of gene symbols and expression data, with the beforementioned user-defined function. The list produced from the output was then unlisted using the sapply() function, and converted into a data frame. The data frame was then written into a file for further used.

**Bootstrap method**

The bootstrap method is done on the statistical programming language R in RStudio. The working directory is first set at the start of the R script. The data frame containing the gene expression data was then read into the R script. The data frame was then reorganised to change the “normal” class columns to the first 4 columns and the “IDC” class columns to the last 16 columns. 2 methods were employed to obtain a binary matrix of significant genes for 1000 resamplings (bootstraps).

Method 1: Using nested ‘for’ loops to produce binary matrix

Method 2: Using “genefilter” package’s “rowttests” function

The resulting output stored in a list was then converted into a matrix by both the unlist() function and the matrix() function. The frequency of significance of each gene out of 1000 resamplings was computed using a ‘for’ loop. The “caret” package was then installed. A confusion matrix was then produced using the confusionMatrix() function from the “caret” package. The precision, recall and F-score were subsequently calculated using the output of the confusionMatrix() function. A visualisation of the confusion matrix was then created by producing a simple plot with a user-defined function. A distance matrix was then computed using a Jaccard’s coefficient user-defined function, and a nested ‘for’ loop. The data frames and matrices were then written into files for further use.

**Results (1 to 3 paragraphs)**

**Discussion (1 to 3 paragraphs)**

**References**

1. Fagard RH, Staessen JA, Thijs L. Advantages and disadvantages of the meta-analysis approach. J Hypertens Suppl. 1996 Sep;14(2):S9-12; discussion S13.
2. Ganesh N. Sharma, Rahul Dave, Jyotsana Sanadya, Piush Sharma, and K. K Sharma. Various Types And Management Of Breast Cancer: An Overview. J Adv Pharm Technol Res. 2010 Apr-Jun; 1(2): 109–126.
3. Higgins JPT, Green S (editors). Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 [updated March 2011]. The Cochrane Collaboration, 2011. Available from <http://handbook.cochrane.org>.
4. Timothy J Key, Pia K Verkasalo, Emily Banks. Epidemiology of breast cancer. The Lancet, Oncology Volume 2, Issue 3, March 2001, Pages 133-140.