Data driven cluster analysis of type 2 diabetes data from Tayside and Fife, Scotland

Anand Thakarakkattil Narayanan Nair[#] , Sushrima Gan[#] ,Adem Y Dawed[#] , Simona Hapca[#] ,Louise Donnelly[#], Ewan R Pearson[#]#School of Medicine, University of Dundee

Introduction

Type 2 Diabetes is a complex, progressive and heterogeneous disease condition. Current type 2 diabetes management guidelines were developed with less attention to the variability in rate of progression and the clinical pathways followed by each diabetic cases . An unsupervised data-driven sub-classification of the adult type 2 diabetes based on the variables assessed at the time of diagnosis may aid in provision of early intensification of treatment for specific groups of individuals. This approach might be useful in predicting the fast progressors in type 2 diabetes continuum at the time of diagnosis and communicating the risk profile will help with more adherence to positive life style changes.

Methods

We analysed recently diagnosed adult type 2 diabetes data (n=29172) from the Scottish Clinical Care Information-Diabetes Care (SCI-Diabetes) to identify unique clusters. Diabetes data from Tayside and Fife region was used in this research. This analysis used five diabetes related variables Age at diagnosis, Body Mass Index (BMI), Glycosylated Haemoglobin (HbA1c) value and High Density Lipid (HDL) levels at diagnosis.

The appropriate number of clusters for the data was determined by adopting “silhouette” and “gap stat” methods. Based on the optimal cluster number we applied a K-Means cluster algorithm on the data after exclusion of outliers.

Cluster validation was done by conducting cluster analysis on partitioned data (test and validating) and sex wise stratified data. Temporal cluster stability was assessed based the longitudinal data on variables selected for cluster analysis (BMI,HbA1c,HDL). Time to first line antidiabetic prescription and second line antidiabetic prescription were assessed by estimating hazard ratios and confidence intervals. Data management and analysis was conducted in “R” version 3.2.5 with RStudio version 0.99.893.

Results

The optimum cluster number for the data was identified as four and cluster analysis identified four replicable clusters. Each cluster showed different characteristics in relation to the five variables included in the analysis. The cluster 1 (Low-Age High-BMI) was 28.0%, cluster 2 (High-Age Low-HDL) was 36.9%, cluster 3 (High-HbA1c) 17.6% and cluster 4 (High-Age High-HDL) was 17.4% of the total study population.

Figure 1: Cluster volume

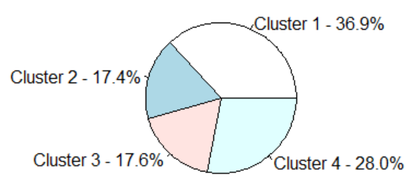
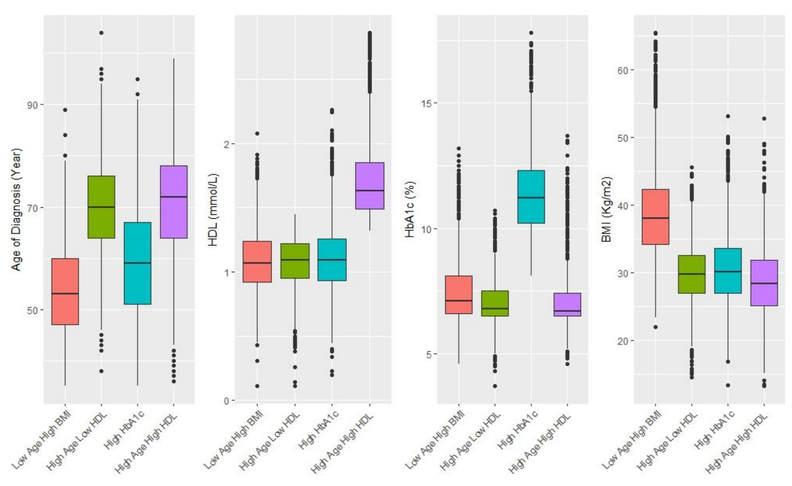


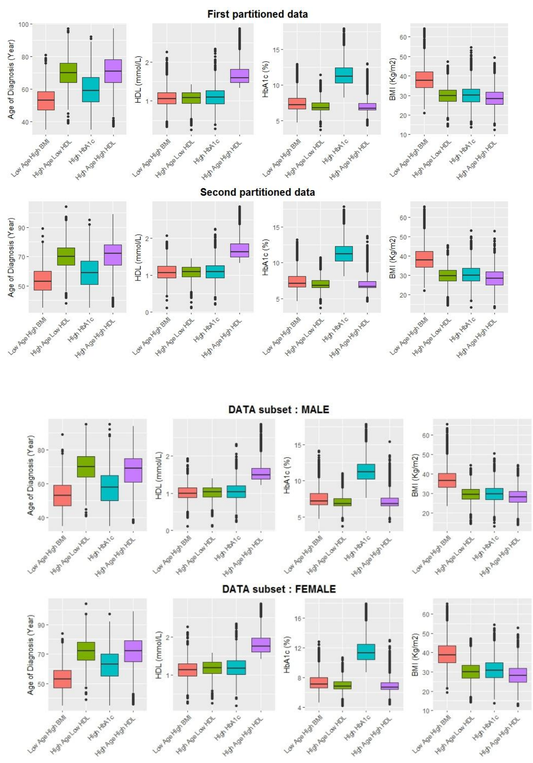
Figure 2: Cluster characterisation



Results

Applying cluster analysis on partitioned (test and validating) data shown similar patterns. Same pattern was visualised across male and female diabetic cases also.

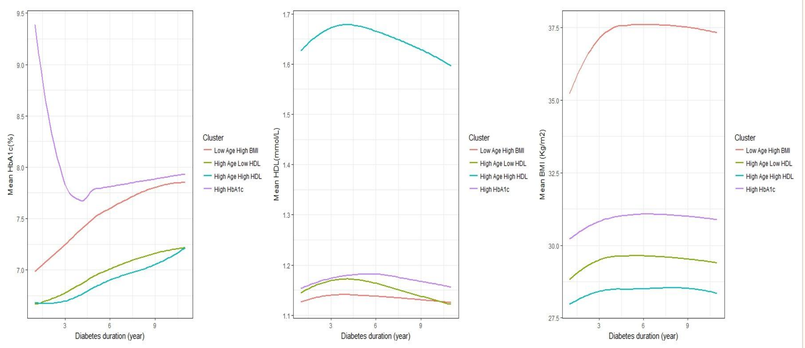
Figure 3: Cluster validation



Results

Four clusters showed stability of their features over 11 year period the time.

Figure 4: Cluster stability (HbA1c,HDL,BMI)



Time to first line *(cluster 3 vs cluster 4; HR 7.72,95% CI 7.28-8.18)* and second line *(cluster 3 vs cluster 4; HR 4.77,95% CI 4.45-5.10)* antidiabetic prescription was shortest in cluster 3 compared to other clusters.

Figure 5 : Time to 1st prescription

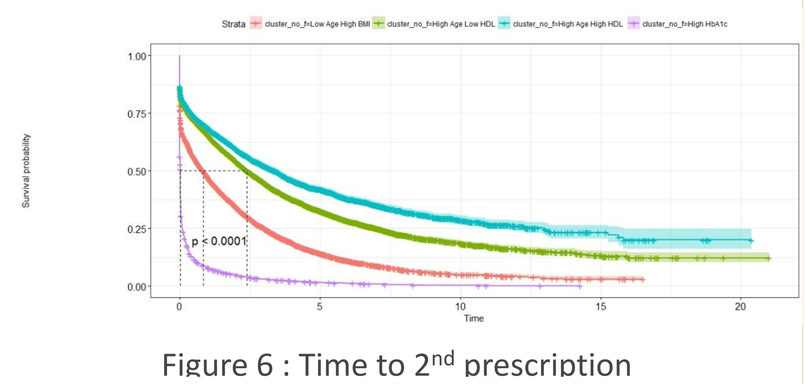
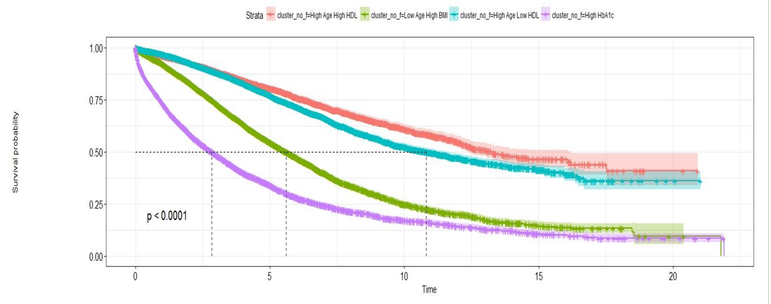


Figure 6 : Time to 2nd prescription



Conclusions

Our analysis identifies four unique type 2 diabetes clusters based on the five variables at the time of diagnosis. Type 2 diabetes cases requiring early intervention can be recognized based on these findings and potentially improving the glycaemic control in this high-risk population

Acknowledgements

Prof. Ewan Pearson’s Research Group, Dr V Mohan, Dr RM Anjana and Health Informatics Centre, Dundee.

This is a summary of independent research funded by the National Institute for Health Research (NIHR)’s INSPIRED Programme. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

Funded by NIHR