A Literature Review Exploring Visual Tools Used for Analysis of Multivariate Data in the Health Sector

GROUP Y

Pablo Quinoa
Anna Agulled Carafi
Kimon Iliopoulos
Konstantinos Gkolias
Yasmin Buzari

1. Analytical Questions

1.1: Analytical questions addressed

The papers reviewed in this discussion were chosen to explore questions in the health and diabetes domain through visual analysis. Key questions understood were how health and socio-economic factors relate to health problems and diseases, such as lung cancer and diabetes, based on practice-level data from NHS care trusts (Borland, West and Hammond, 2014). Another paper questioned possible trends that can be found in high volume patient data to give insight into the pathogenesis of diseases (Kolesnichenko et al., 2019). Within the context of diabetes, analysis methods were more focused on prevalent factors such as obesity, parental history, and genetics to the contribution of developing type 2 diabetes (T2D) (Berumen et al., 2019). In relation, observing differences in parallel of healthy versus T2D patients distinguished by levels of trace elements and blood samples were key aim's that was explored through visual clustering methods (Badran et al., 2016) (Sitnikova et al., 2018). Thus, the domain chosen was specific to understanding key questions in improving health factors and prevalence of diabetes. Its specificity indicates that there is difficulty in applying these questions to other domains and contexts', because of the data types of health variables.

1.2: In which particular contexts is the type of data being studied

A generalised perspective on health and pathology of diabetes was taken to study multivariate data.

2. Aspects of the data and phenomenon

2.1: Important issues when interpreting or analysing data types

Within the health domain, a common challenge found is the mix of different data types, such as categorical, numerical discrete, numerical continuous being analysed in combination to observe trends or findings (Borland, West and Hammond, 2014). Another aspect understood from our observations is the difficulty to isolate factors or variables contributing to specific health problems, such as lung cancer. For example; a doctor cannot confirm the diagnosis of lung cancer based on a single factor such as smoking but a combination of factors. In addition, due to high volume of data, there is a challenge in extracting useful patterns or trends when there is noise in data and/or unwanted variables that would not be purposeful (Gamberger, Lavrac and Dzeroski, 2000). Data format is also another challenge because most patient management data is operational and therefore difficult to pre-process for use in analysis (Kolesnichenko et al., 2019).

2.2: What is unique about the chosen domain?

The chosen domain is unique because of the challenge in visually portraying the datasets studied in the papers. Additionally, another aspect is the high dependency of factors contributing to specific health problems such as lung cancer, which can prove difficulty in finding single dependency (one dependant in causation to one independent factor), thus multivariate representation becomes necessary.

3. Structured comparison

By measuring 26 variables in health such as diabetes prevalence or cancer rates, and socio-economic factors like geographic region or socio-economic deprivation we can infer relations between these factors (Borland, West and Hammond, 2014). Because of the large number of variables being analysed, specific methods such as **radial coordinate visualisations** where each variable is an axis of the radial, is preferred.

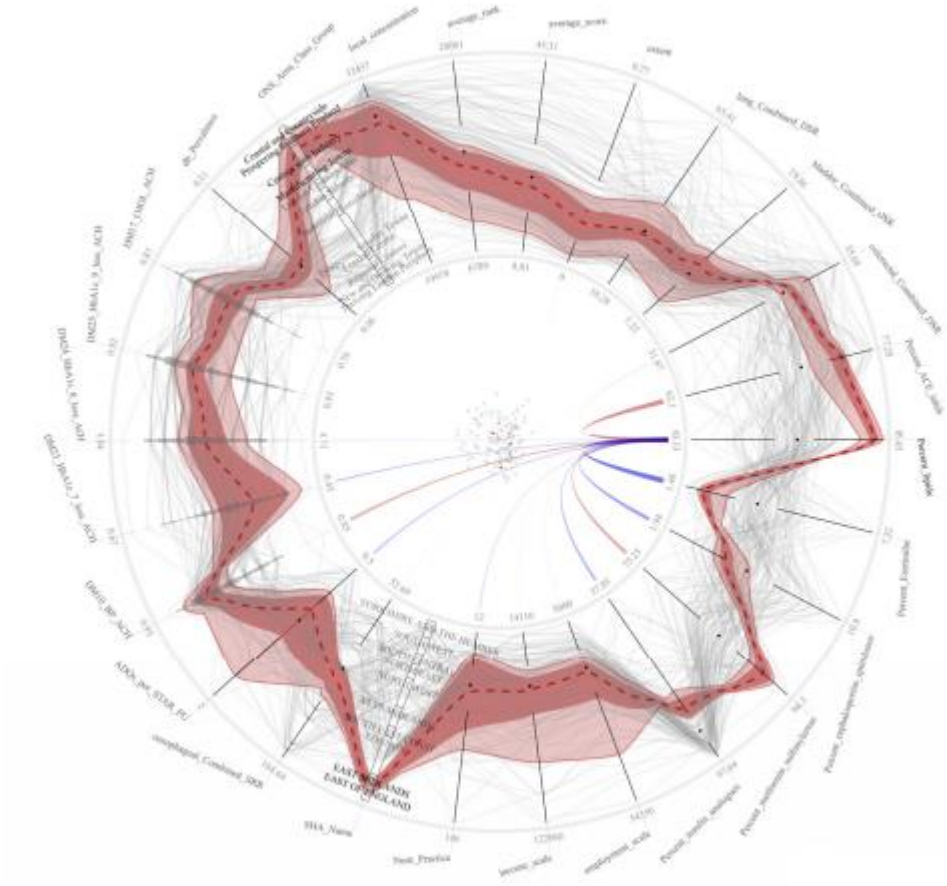

Figure 1. Radial Coordinates

Curves in grey and red represent PCTs (different Primary Care Trusts across the UK), so we can see at what level each curve intersects each of the 26 axes. By selecting an individual axis (variable), correlation with the rest of the variables can be seen with red and blue curves (smaller internal circle Figure 1), where red indicates a positive correlation and blue a negative correlation (Borland, West and Hammond, 2014). An axis can be of different data types - numerical discrete or continuous, and categorical, so for continuous numeric axes as a plot divided into quartiles, a histogram representing discrete numeric, and a stacked bar chart representing the categorical axes is shown in Figure 2 (Borland, West and Hammond, 2014).

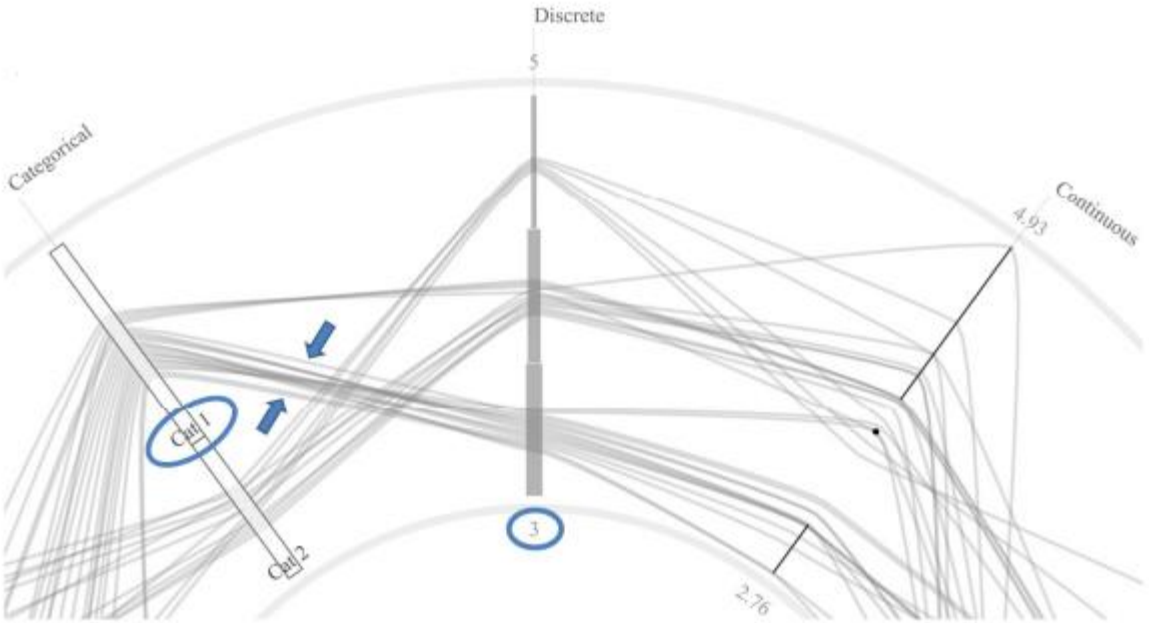

Figure 2. A zoom in to the 3 different data types axes from the radial coordinates visualisation

Therefore, radial coordinates accommodate very well to compare multiple health related variables of different data types in just a single graph, to then be able to infer causes for different health problems.

Another method of analysis by visualisation to infer trends across high volume data is by using **radial cluster plot**. Cluster analysis focused on measures of haemoglobin glytate and parathyroid-hormone associated with specific medical procedures forming cluster groups (Kolesnichenko et al., 2019).

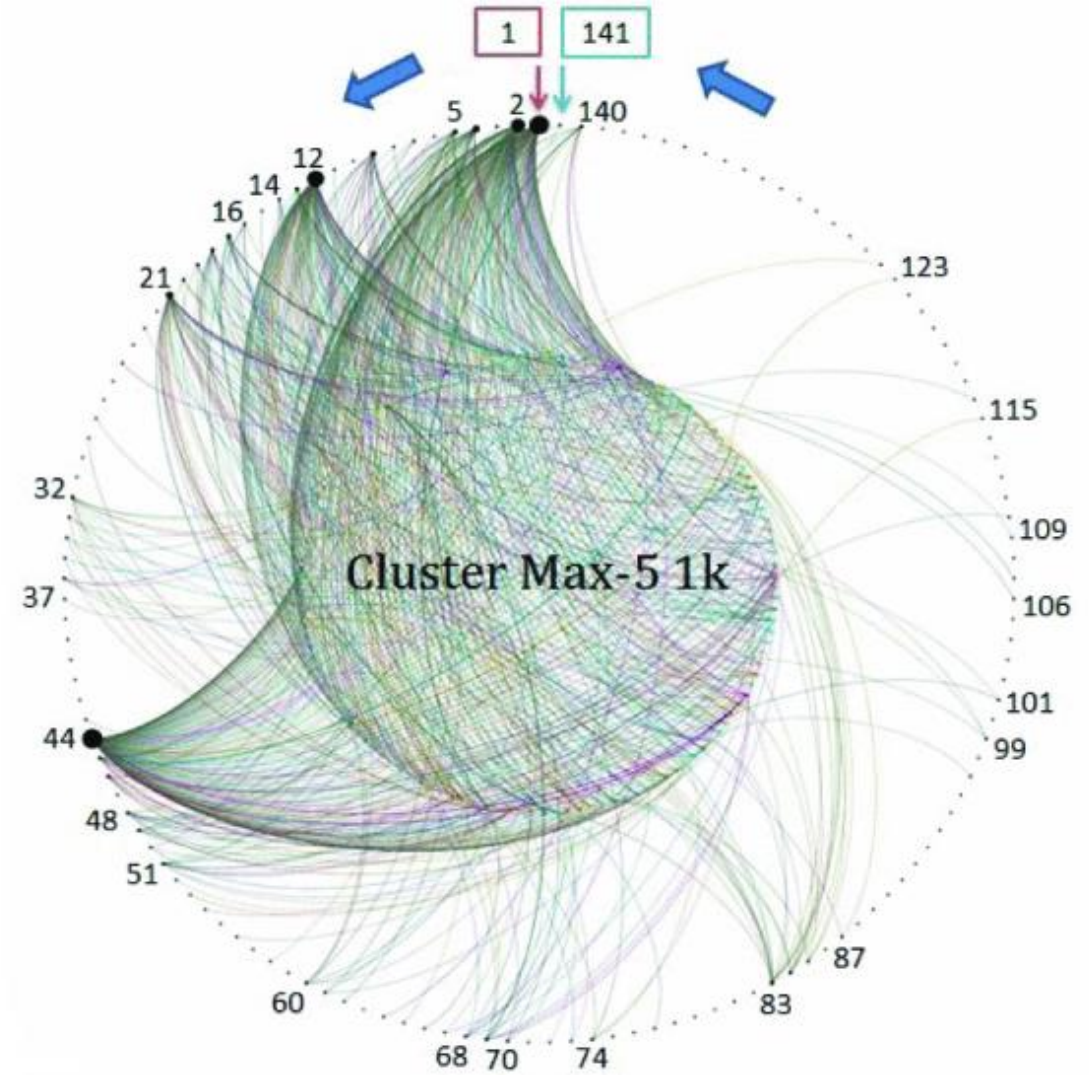

Figure 3. Radial Cluster Plot

The centroid of Figure 3. plots patients diagnosed with type 1 diabetes (T1D) matched to black dots representing tests/surgeries mapped numerically on the edge so for example; point 44 represents lung x-ray referred from the meta-data table in the study (Kolesnichenko et al., 2019). We can infer from this that there is a correlation with T1D and a likelihood onset of lung cancer, because of the convergence of patients administered to lung x-rays. Thus, the first visualisation by Borland, West and Hammond, 2014 is very useful to see how variables of multiple data types correlate with each other, whilst the cluster graph highlights trends (concentration of T1D patients to a procedure).

Another paper showed the use of hierarchical cluster analysis, and this is useful to visualise a binary outcome (healthy vs. T2D patients). Figure 4. **Dendrogram** is a suitable method of visualisation for health, because of the link in chemical associations with patients (Badran et al., 2016).

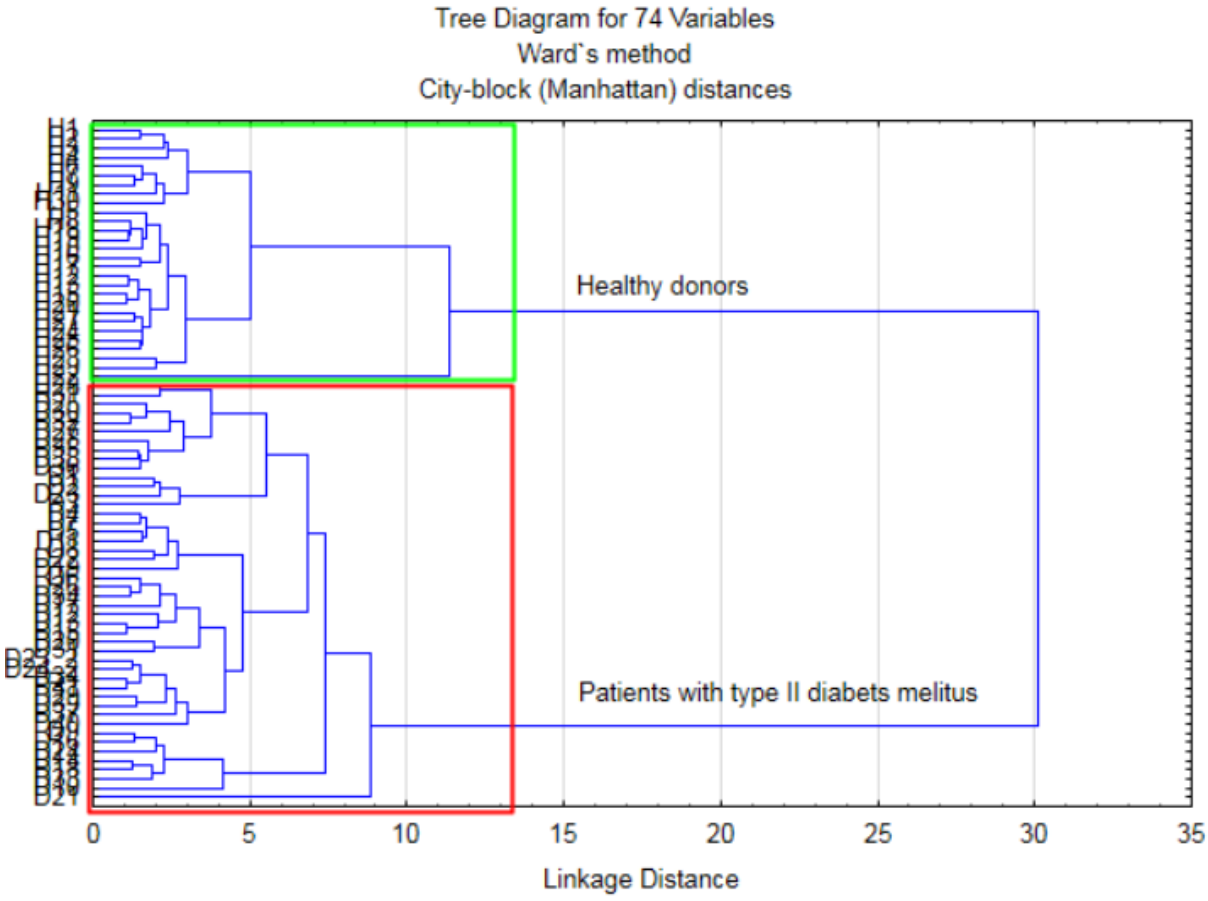

Figure 4. Dendrogram

Another research method helps us visually identify the role of genes WFS1 and INS-IGF2 as deterministic variables, by plotting the results of the multivariate logistic regression for males (blue), females (red), and both genders combined (red) as seen in Figure 5 (Berumen et al., 2019). A limitation to this method is its suitability only in specific comparisons, thus producing multiple visualisations. This is also a common problem in multivariate graphical representation, which can be improved by merging more than two variables in one graph like in the radial visualisations.

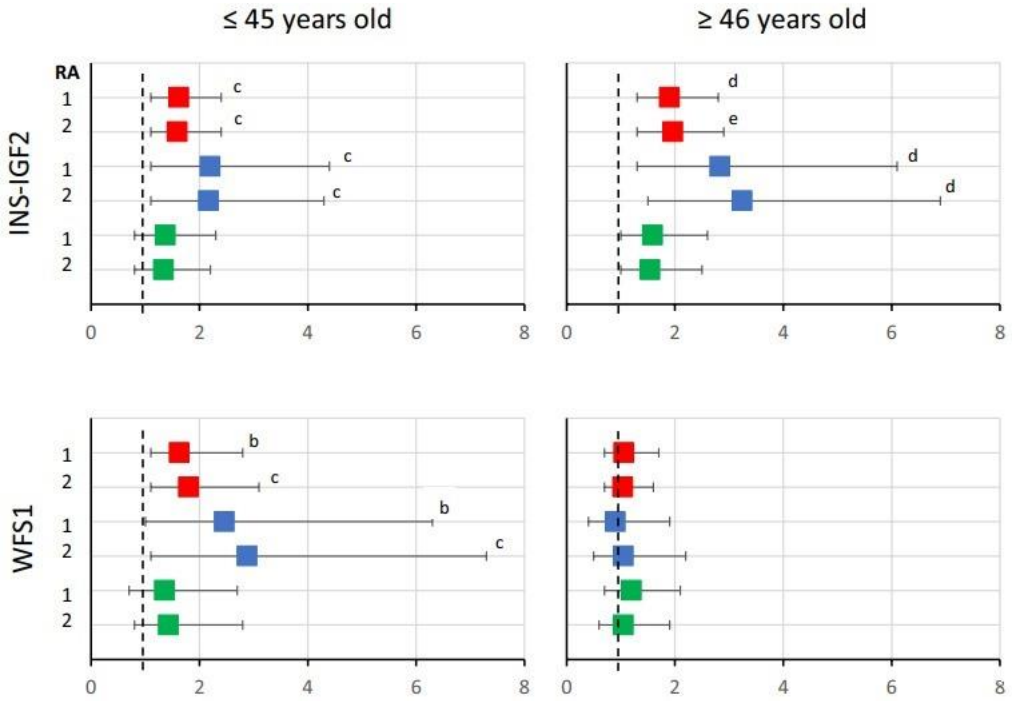

Figure 5. Odd Ratio Plot

In conclusion, the findings shown through visually analysing the data in the examples presented here can be used to explore vital questions further in the health domain.

References

Badran, M., Morsy, R., Soliman, H. and Elnimr, T. (2016). Assessment of trace elements levels in patients with Type 2 diabetes using multivariate statistical analysis. *Journal of Trace Elements in Medicine and Biology*, [online] 33, pp.114-119. Available at: https://0-www-sciencedirect-com.wam.city.ac.uk/science/article/pii/S0946672X15300419 [Accessed 16 Oct. 2019].

Berumen, J., Orozco, L., Betancourt-Cravioto, M., Gallardo, H., Zulueta, M., Mendizabal, L., Simon, L., Benuto, R., Ramírez-Campos, E., Marin, M., Juárez, E., García-Ortiz, H., Martínez-Hernández, A., Venegas-Vega, C., Peralta-Romero, J., Cruz, M. and Tapia-Conyer, R. (2019). Influence of obesity, parental history of diabetes, and genes in type 2 diabetes: A case-control study. *Scientific Reports*, [online] 9(1), pp.1-15. Available at: https://www.nature.com/articles/s41598-019-39145-x.pdf [Accessed 7 Oct. 2019].

Borland, D., West, V. and Hammond, E. (2014). Multivariate Visualization of System-Wide National Health Service Data Using Radial Coordinates. *Visual Analytics in Healthcare*, [online] pp.53-59. Available at: https://www.visualanalyticshealthcare.org/docs/VAHC2014_proceedings.pdf [Accessed 12 Oct. 2019].

Gamberger, D., Lavrac, N. and Dzeroski, S. (2000). Noise detection and elimination in data preprocessing: Experiments in medical domains. *Applied Artificial Intelligence*, [online] 14(2), pp.205-223. Available at: https://www.tandfonline.com/doi/abs/10.1080/088395100117124 [Accessed 21 Oct. 2019].

Kolesnichenko, O., Marochkina, E., Komarov, R., Mazelis, L., Mazelis, A., Soldatov, D., Minushkina, L., Chernoskutov, M., Averbukh, V., Mikhaylov, I., Martynov, A., Pulit, V., Amelkin, S., Grigorevsky, I. and Kolesnichenko, Y. (2019). Big Data Analytics of Inpatients Flow with Diabetes Mellitus type 1 : Revealing new awareness with Advanced Visualization of Medical Information System Data. *2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*. [online] Available at: https://0-ieeexplore-ieee-org.wam.city.ac.uk/document/8776910/authors [Accessed 16 Oct. 2019].

Sitnikova, V., Nosenko, T., Olekhnovich, R. and Uspenskaya, M. (2018). Multivariate Analysis for Diagnostic of Type II Diabetes Mellitus. *2018 IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES)*. [online] Available at: https://ieeexplore.ieee.org/document/8626635 [Accessed 21 Oct. 2019].