Exploratory data analysis on the BT Life Sciences Workbench: Can the higher incidence of diabetes among South Asians in England be validated with available data?

Abstract

Visually inferred correlations biased by prior knowledge subsequently require a rigorous validation. An example is the increased propensity for South Asians to develop diabetes. Our work points to a common danger in biased inference arising from data aggregation. Validation tools used in exploratory data analysis are also described.

The 2015 BT/SRII Health Data Challenge

The 2015 Health Data Analytics Challenge was sponsored by British Telecom in the Americas (BT) and the Service Research and Innovation Institute (SRII). Participating teams were encouraged to formulate and investigate questions of interest utilizing the cloud-based life sciences workbench provided by BT along with public health data from the UK National Health Service (NHS) and other sources.

NHS public health data is available at various levels of aggregation, such as the Strategic Health Authorities, Primary Care Trusts (PCTs; also referred to here as districts), hospitals, social care, mental health and individual practices. Data at the practice level includes age groups, prescription details and prevalence of certain diseases, as specified by the NHS Quality and Outcomes Framework (QOF). However, de-identified patient-level data is not available.

Starting Hypothesis: Available data confirms that UK residents of South Asian origin have a greater likelihood of developing diabetes compared to those of European origin

A multitude of studies have established that populations originating from South Asia (Bangladesh, India, Pakistan) have an increased propensity to develop diabetes (see the references in [1]). Establishing a direct dependence of diabetes prevalence at the practice level on the share of patients of South Asian origin could justify a targeted patient outreach strategy motivating diet and lifestyle changes [2].

Does an overlay of two disparate datasets confirm the starting hypothesis?

While NHS diabetes prevalence data for England is available at the practice level (practice_qof_prev_1011), ethnicity data is only available at the district level (adm_census_ethnic_mix).

	Reported QOF	Total
Practices	8246	10098

Table 1: Comparison of the number of total practices (practice_add) and the number of practices that reported their QOF (practice_qof_prev_1011).

Figure 1 depicts an overlay of the two datasets for the country of England, where district-level coordinates have been retrieved from Google maps (ggmap package in R). At first glance, this visual mapping could suggest a correlation between the South Asian population and the prevalence of diabetes.

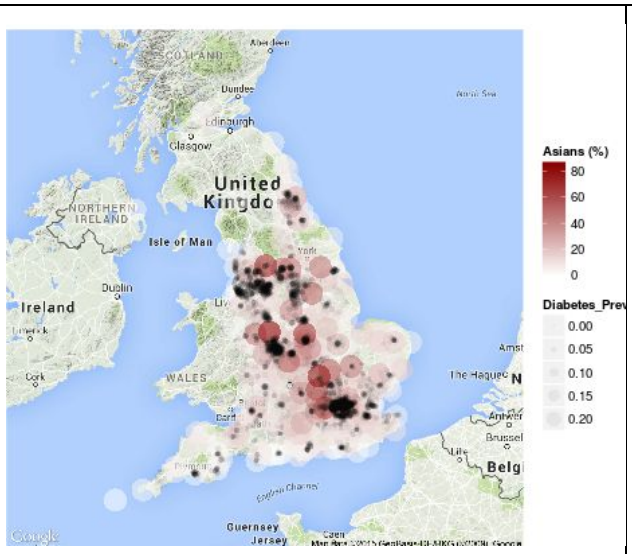


Figure 1: Map of England with South Asian population percentage at the district level represented by shaded circles, and the prevalence of diabetes by size of the black dots.

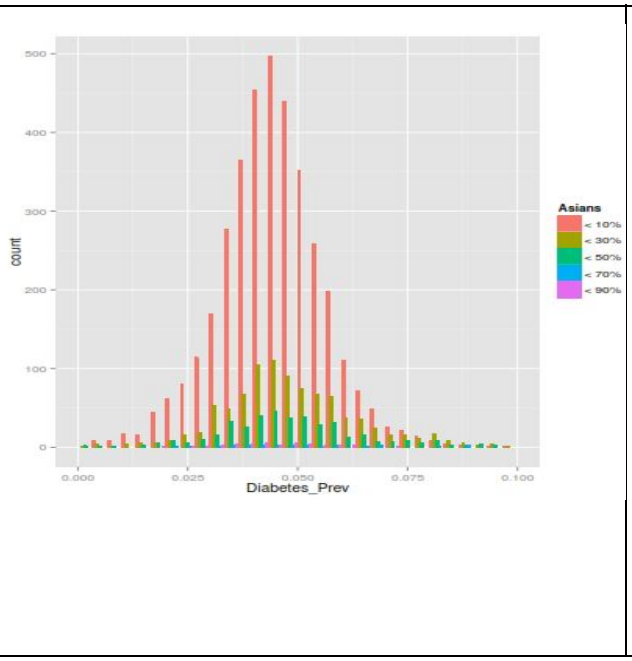
Population data is at the district level, while prevalence data comes from reporting clinics.

Does the overlay map indicate a genuine correlation between diabetes incidence and the percentage of South Asian population?

As the ethnic mix of patients at the practice level is not available, we approximated it using the overall value for each district (or PCT). Using the postcode locations (obtained from freemaptools.com), each practice was assigned to the nearest district center. We recognized that this assumption would smoothen out ethnic variations between individual practices.

For further analysis, each district was assigned to one of five demographic categories based on its percentage of South Asian population, .

Does the category representing the smallest percentage of South Asians also show the least prevalence of diabetes? Figure 2 provides both a histogram comparison and a boxplot representation of each of the five categories. Distributions of diabetes prevalence appear similar in each category, with the peaks (distribution medians) and spread appearing comparable.



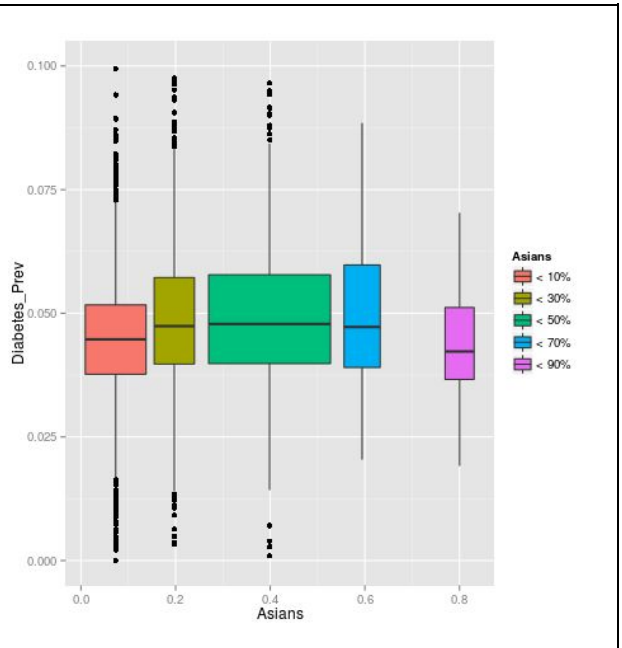


Figure 2: Diabetes prevalence was aggregated for districts within five specific demographic categories representing ranges of the percent population of South Asians (with midpoints at 10%, 20%. 40%, 60% and 80%). Boxplots reveal little variation in the distributions.

Hypothesis rejected

A t-test detected no significant difference

A t-test validated the observations from our boxplot analysis. It revealed no significant difference among the various demographic categories representing percentages of South Asians in districts across England.

The overlay map in Figure 1 therefore presents an artifact, which is attributable to two sources. The first source is the relatively small proportion (5.7%) of South Asians across England. Figure 1 indicates that they are dispersed over a wide area. The next possible source of error is the relatively large number of practices that did not report on QOF measures.

Principal Component Analysis (PCA) did not find diabetes prominent among the first two components of variance of all morbidities or health risks

While the prevalence of diabetes appeared relatively uniform in all demographic categories, we asked whether this was the case for all morbidities and risk factors. Accordingly, we conducted principal component analysis (PCA) on the practice-level QOF data to see where the majority of the variance lies? PCA is a tool to determine the principal directions of variance in a dataset comprising many variables. From Figure 3 we see that the first component is predominantly heart failure due to left-ventricular dysfunction (accounting for 26% of the variance) and the second component is a mix of smoking indicators and hypertension (6% variance). While this does not cover a large degree of variance, we see that diabetes is not the disease that contributes to the majority of the data.

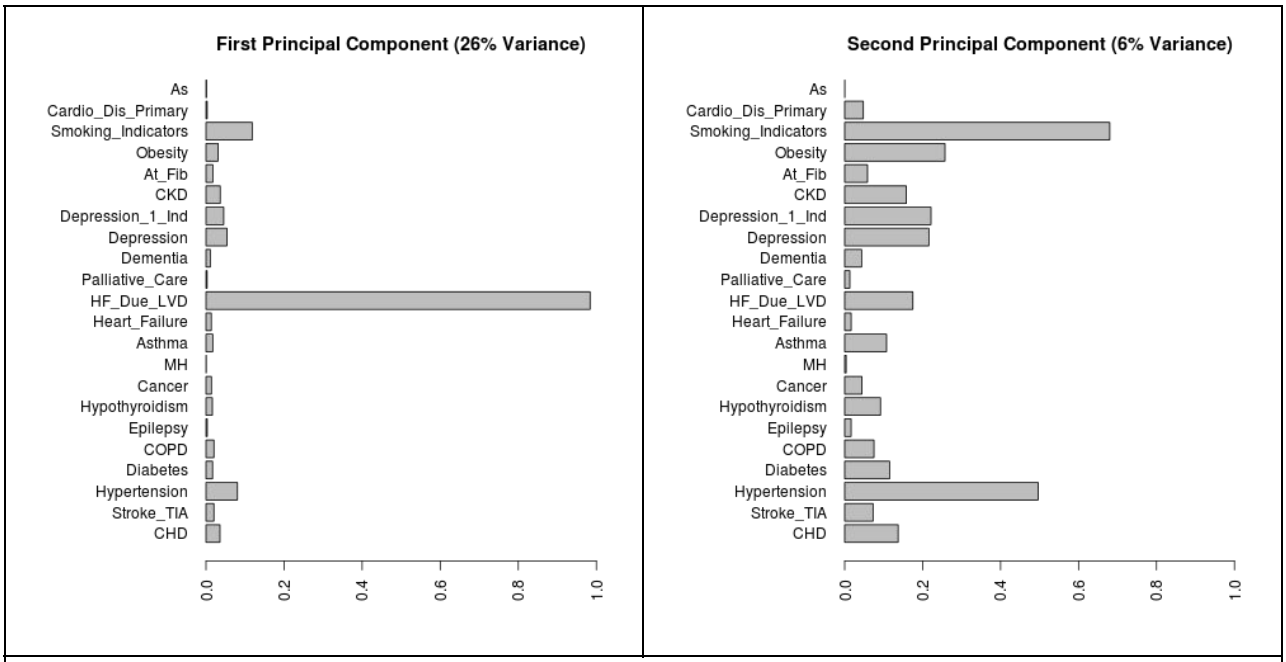


Figure 3: The first (left) and second (right) principal component of the prevalence and ethnic data at the practice level.

Conclusion

Preliminary data analysis can often mislead us into seeing spurious correlations, especially given biases based on previous knowledge. The case in point is the propensity of South Asians in England to develop diabetes (as conveyed in Fig. 1). Statistical analysis using boxplots and t-tests reveals that such a propensity is masked out in aggregate data. Ideally, the more granular de-identified patient-level data could ensure a more thorough analysis of the well-documented claim [1]. However, it would be interesting to also validate such a correlation in practice-level data where the ethnic composition of the patient population is available.

References

[1] Kamlesh Khunti, Sudhesh Kumar, Jo Brodie: *Diabetes UK and South Asian Health Foundation recommendations on diabetes research priorities for British South Asians,* ***Diabetes UK****, 2009*

[2] Carlos A. Celis-Morales, Nazim Ghouri, Mark E. S. Bailey, Naveed Sattar, Jason M. R. Gill: *Should Physical Activity Recommendations Be Ethnicity-Specific? Evidence from a Cross-Sectional Study of South Asian and European Men,* **PLOS One**, Vol. 8, Issue 12, 2013

Methods/Toolbox

The data analysis in this report has been done in R. The figures have been made using the packages of ggmap (Google maps, Fig. 1) and ggplot2 (Grammar of graphics, Fig. 2 and 3).

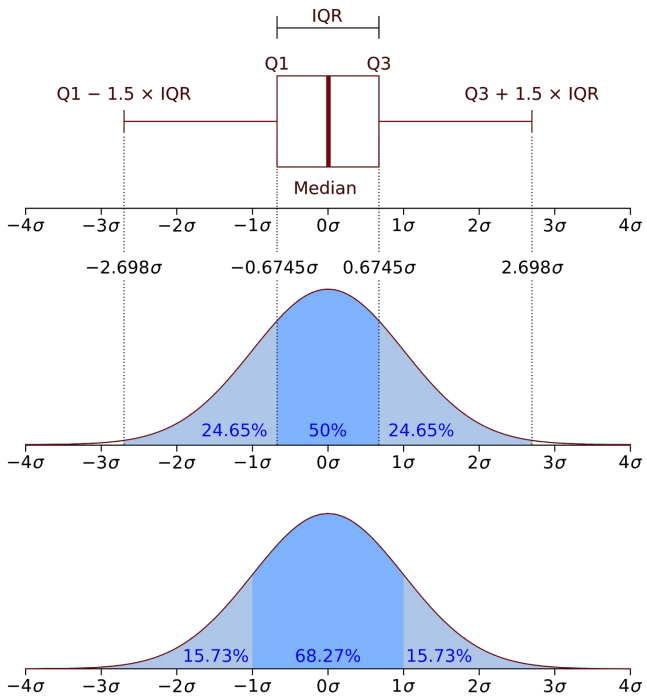
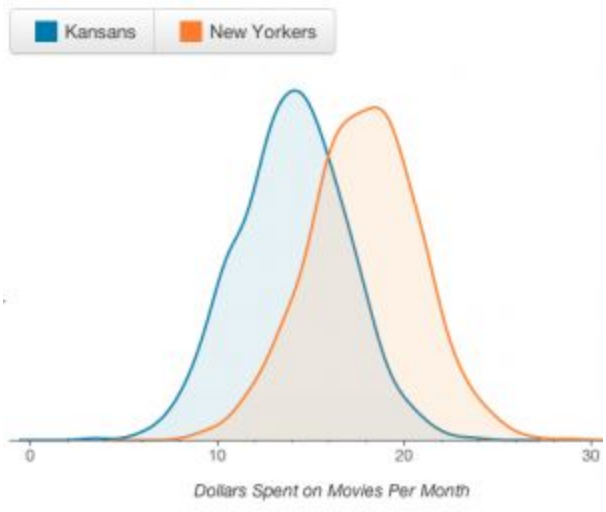


Figure 4: Depiction of the boxplot in reference to the normal distribution. Note that the 1st and 3rd quartiles (Q1 and Q3) do not correspond to the standard deviation (sigma). [Image from Wikipedia]



Given two sample groups with different average values, a t-test determines how likely they are to be drawn from the same population. For example, whether people from Kansas spend less money on movies per month compared to people from New York (see Fig. 5), by asking if both samples are likely to be from the same parent population of movie goers.

Figure 5: Overlay of the distributions of dollars spent on movies for Kansans (blue) and New Yorkers (orange). [Image from statwing.com]

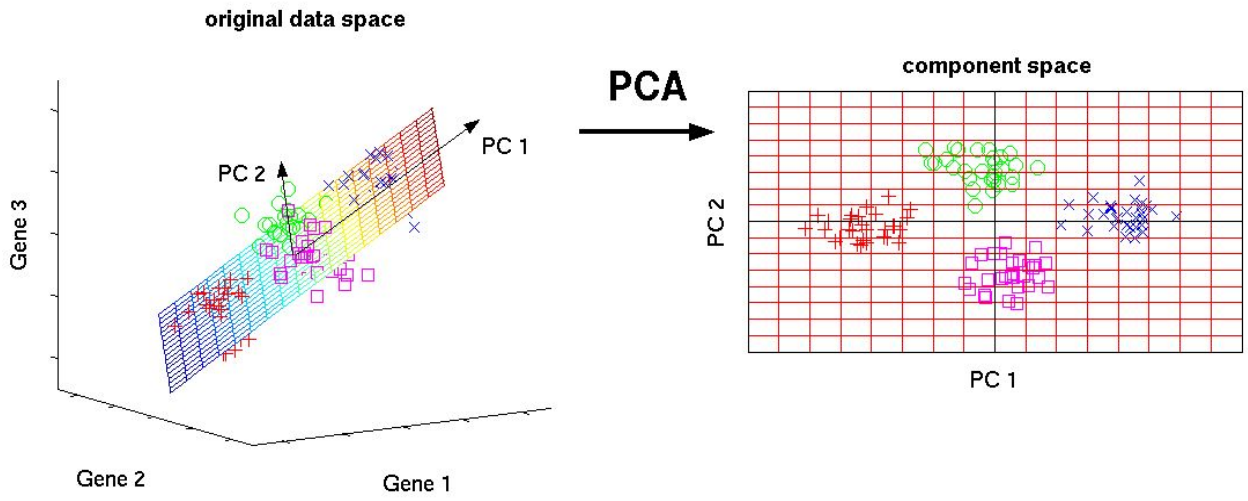


Figure 6: PCA finds a set of axes (principal components) describing a distribution of several variables. On the left we see data points plotted in three dimensions. By looking at just the first two principal components, we can project the data onto two dimensions (right), from which we can distinguish the different classes associated with the data. [Image from nlpca.com]