PRM66

Synthetic Sample Generation Representing the English Population using Spearman Rank Correlation and Cholesky Decomposition

Cassie Springate;[1] Chris Martin[1]

[1] Crystallise Ltd cassie.springate@crystallise.com, chris.martin@crystallise.com

Objectives:

To generate a synthetic sample of 1 million individuals that reflects the characteristics of the population recorded in the Health Survey for England.

Methods:

- We used data from the Health Survey for England (HSE) to determine the age- and gender-dependent distributions of continuous variable risk factors **(height, weight, BMI, systolic blood pressure, total and HDL cholesterol and their ratio, number of cigarettes/day and units of alcohol/week)** and prevalence of binary risk factors **(smoking status, diabetes)**.
- Spearman rank correlations including age and gender were determined for the risk factors and a table of normally distributed random numbers was generated. Cholesky decomposition was used to replicate the observed Spearman rank correlations in the table of random numbers.
- Rank correlations that included binary variables were recalibrated to adjust for numerous tied values.
- The synthetic sample (SS) was generated using a reverse look-up of the gamma distribution value using the random percentiles for continuous variables or setting a binary variable to 1 when the random percentile falls below the prevalence threshold.

Results:

- Differences between coefficients were no more than 0.5% for any continuous variable.
- The prevalence of binary factors in the SS was very well matched with the HSE sample.
- Comparing 25th, 50th and 75th quantiles, the maximum difference between the original and synthetic values for BMI and TC/HDL ratio were 0.6kg/m^2 and 0.3 respectively.

Figure 1: Prevalence of binary risk factors as a function of gender and data sample

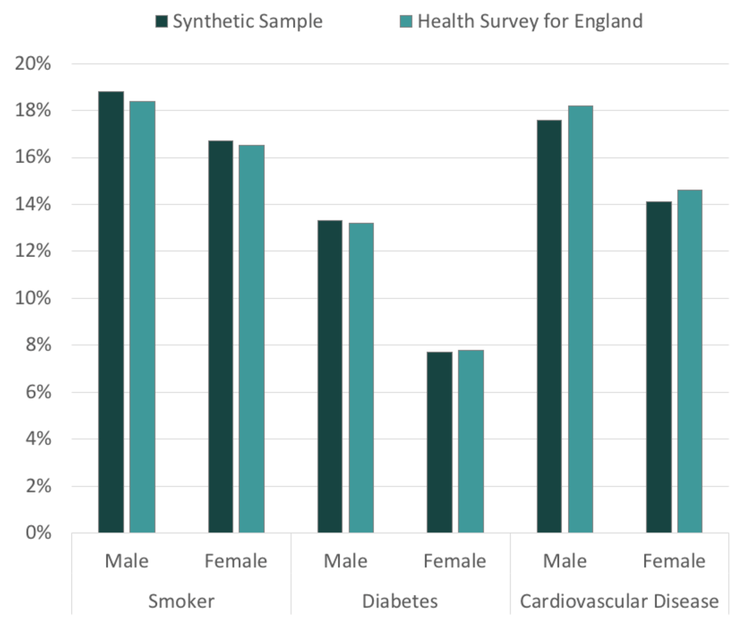
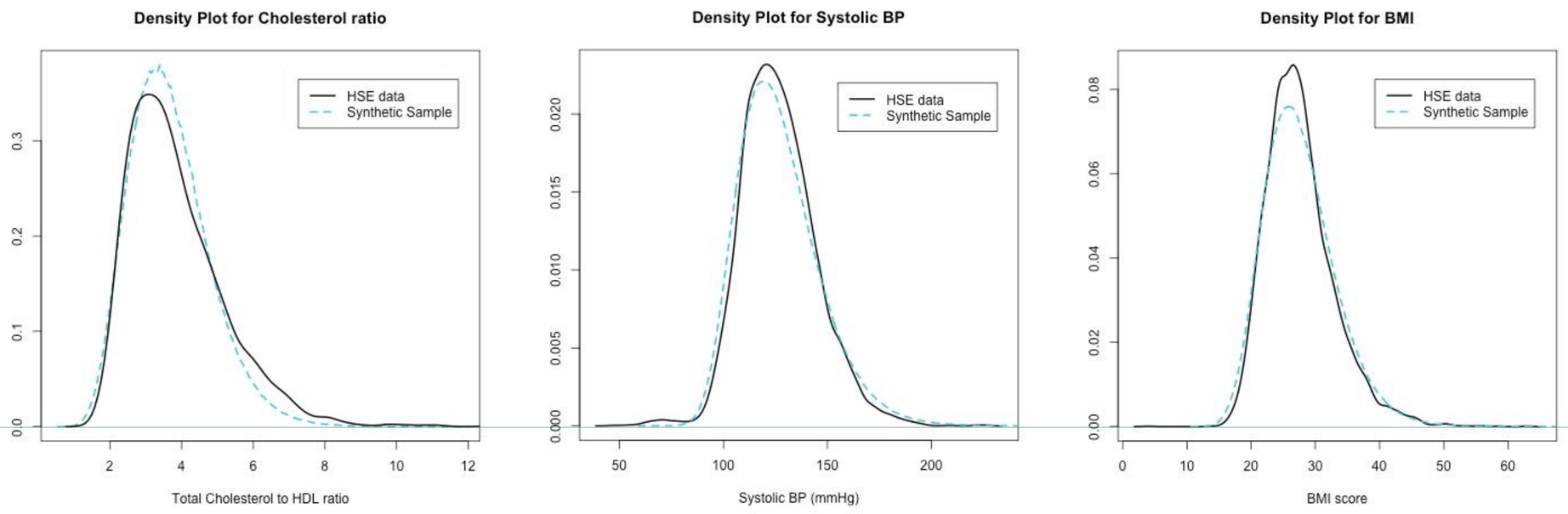


Table 1: Absolute differences between partial correlation coefficients for the HSE data compared with the synthetic sample

	Height	Weight	BMI	SYSTOLIC	CPD	Alcohol units/week	Total Cholesterol	HDL	TC to HDL ratio	Diabetes	CVD
Height	.00	.01	.02	.00	.00	.00	.01	.00	.03	.00	.00
Weight	.01	.00	.01	.02	.02	.01	.02	.02	.02	.01	.01
BMI	.02	.01	.00	.03	.03	.01	.01	.05	.04	.01	.00
SYSTOLIC	.00	.02	.03	.00	.01	.01	.02	.02	.00	.01	.00
CPD	.00	.02	.03	.01	.00	.01	.01	.03	.02	.01	.01
AlcUPW	.00	.01	.01	.01	.01	.00	.01	.01	.02	.01	.01
TC	.01	.02	.01	.02	.01	.01	.00	.00	.05	.02	.02
HDL	.00	.02	.05	.02	.03	.01	.00	.00	.03	.03	.00
TC2HDL	.03	.02	.04	.00	.02	.02	.05	.03	.00	.00	.01
DM	.00	.01	.01	.01	.01	.01	.02	.03	.00	.00	.01
CVD	.00	.01	.00	.00	.01	.01	.02	.00	.01	.01	.00

Figure 2: Density plots showing the original HSE sample data against the generated synthetic sample



Conclusions:

- Our new approach generates large synthetic samples with risk factor distributions very closely matching those of the real HSE population.
- This sample can be used to model the likely impact of new therapies or predict mortality.

Scan for supplementary data:

Crystallise Ltd. Unit 21 Thames Enterprise Centre, Thames Industrial Park, East Tilbury, Essex UK RM18 8RH **Tel**: +44 01375 488020

For a copy of this poster, email: chris.martin@crystallise.com

www.crystallise.com www.heoro.com

Presented at the ISPOR 23rd Annual International Meeting May 19-23 2018; Baltimore, MD, USA