PRM114

Synthetic Sample Generation of the 4S Study Placebo Population Using a Stochastic Sampling Technique

Chris Martin[1]; Cassie Springate[1]
1 Crystallise Ltd chris.martin@crystallise.com, cassie.springate@crystallise.com

Objectives:

- To generate a synthetic sample of individuals with similar average values to the risk factor variables in the 4S study sample, and similar distributions and correlation structure compared with the general population.

Methods:

- A base synthetic population sample was used that captures the same distribution of risk factor variables and correlation structure as the Health Survey for England, as described in a previous publication.[1]
- **A stochastic resampling technique** was used to generate a semi-random sample of people with characteristics that match those of the control group in the **Scandinavian Simvastatin Survival Study (4S)**, using R.
- The sample was matched on:
  - binary variables; gender, smoking status, diabetes
  - continuous factors; age, BMI, systolic blood pressure, total cholesterol: HDL cholesterol ratio, cigarettes smoked per day and units of alcohol per week
- The mean values for the risk factors matched the target sample to an accuracy of 1 decimal point.

Results:

The algorithm successfully generated a sample of 2,222 individuals with characteristics closely matching those of the 4S study control group.

- The 4S study only reported descriptive statistics for patient characteristics.
  - For distributions we compared the generated sample data with the Health Survey for England 2012 population data.

Descriptive statistics:

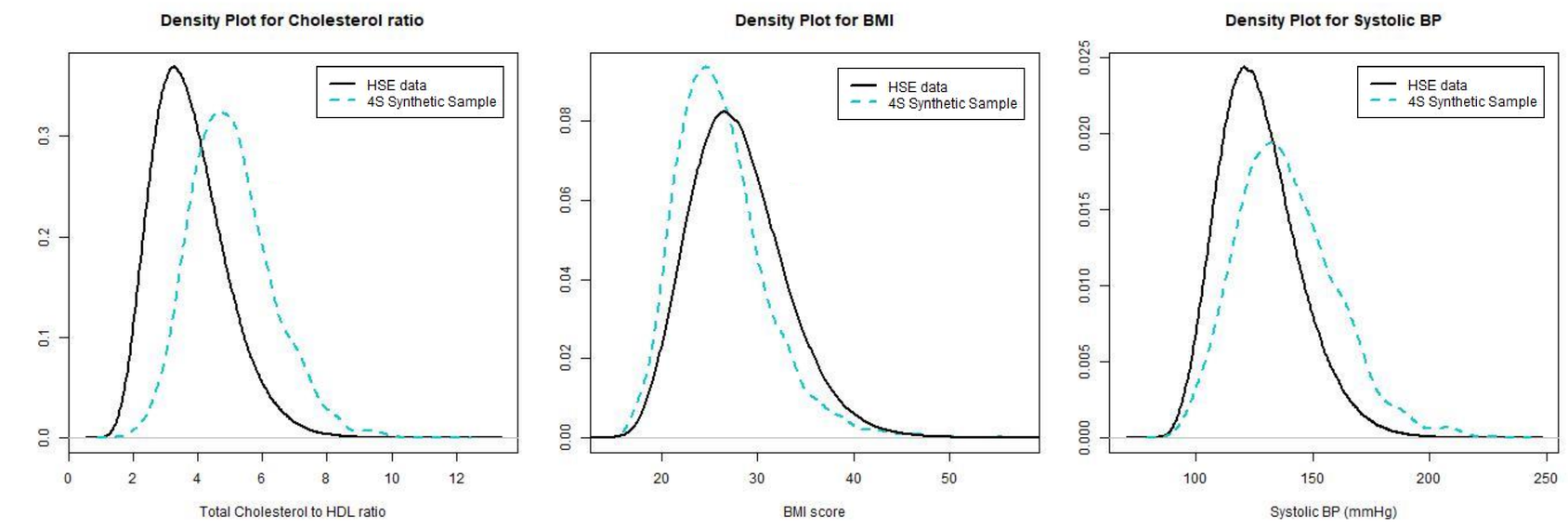
	Synthetic Sample	Original 4S Sample	% Error
Male age	58.1	58.1	0%
Female age	60.51	60.50	0%
Systolic BP	139.1	139.1	0%
TC	6.75	6.75	0%
HDL	1.19	1.19	0%
BMI	26.0	26.0	0%
Males	81%	81%	0%
Females	19%	19%	0%
Smokers	25.3%	25%	1%
Diabetes	4.3%	4.5%	4%

Synthetic sample correlations: (blank cells indicate a statistically non-significant correlation)

	AGE	SEX	Cigarettes Per Day	SYSTOLIC	BMI	TC / HDL	Alcohol Units/Week
AGE	NA	-10%	-13%	29%	18%		
SEX	-10%	NA	6%		7%	21%	14%
Cigarettes Per Day	-13%	6%	NA	-7%		-8%	
SYSTOLIC	29%		-7%	NA	38%	9%	
BMI	18%	7%		38%	NA	37%	
TC / HDL		21%	-8%	9%	37%	NA	4%
Alcohol Units/Week		14%				4%	NA

- As expected, there is a high correlation between systolic blood pressure and BMI.
- There is also a high correlation between BMI and total cholesterol / HDL ratio.

Density Plots: Synthetic Sample data compared with real world HSE population data



- As expected, the distribution of total cholesterol / HDL ratio, BMI and systolic blood pressure were skewed upwards reflecting the higher average values in the high-risk 4S sample.

Conclusions:

- We successfully generated synthetic samples that are comparable to the originals in aggregate.
- Our approach can be used to model the likely impact of new therapies or predict mortality for various sub-groups.
- This will be a useful tool in the planning and preparation of clinical trials.

Refences: [1] Martin, C., & Springate, C.E. Synthetic Sample Generation Representing the English Population Using Spearman Rank Correlation and Chomsky Decomposition. Presented at ISPOR Baltimore 2018, PRM66

Scan for online pdf:

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 ISPOR Europe 2018
November 10th-14th; Barcelona, Spain