Suppl_imputation.pdf

Rapsomaniki

Prognostic models for stable coronary artery disease based on electronic health record cohort of 102,023 patients

Multiple imputation

The R package 'mice'[1], version 2.0 was used for imputation. The following variables were imputed: smoking status, ethnicity, BMI, systolic and diastolic blood pressure, heart rate, total cholesterol, low-density and high-density lipoprotein, white cell count, creatinine, haemoglobin, Hba1c, gamma glutamyl-transferase (GGT) and unclassified MIs. Among these variables, GGT, Hba1c and heart rate were recorded the least frequently (14-22%), creatinine the most frequently (67%) and the rest had approximately 45-50% missing data. Missing data was less frequent in patients with more comorbidities or those on medications (e.g. patients on statins were 3 times more likely to have their cholesterol recorded and patients with renal failure were more likely to have creatinine recorded). After multivariate adjustment data appeared to be missing at random, except for Hba1c that was almost exclusively measured in patients diagnosed with diabetes (hence not considered further).

Each risk factor was imputed using all other covariates as predictors, as well as their average levels in the year prior to the index diagnosis date (where relevant). Imputation models included the event indicator and the Nelson-Aalen baseline hazard to t, where t denotes the time to event/censoring for each patient[2]. Imputation was based on logistic regression for binary variables and Bayesian linear regression[3] for continuous variables, after log-transformation of non-Normally distributed variables. Models were fitted in 20 multiply imputed datasets and the coefficients combined following Rubin's rules[4]. Missing data patterns were assessed by comparing observed risk among patients with and without the variable of interest recorded. In the following table we compare age and sex adjusted hazard ratios between imputed and complete cases.

Observed and imputed hazard ratios adjusted for age and sex for variables with missing data included in the CALIBER models for mortality and non-fatal MI or coronary death risk.

	All-cause mortality			Non-fatal MI or coronary death	
Variable	Complete cases %	Imputed HR [95% CI]	Complete casesHR [95% CI]	Imputed HR [95% CI]	Complete cases HR [95% CI]
Current smoker vs. never	68	1.33 (1.27-1.38)	1.45 (1.39-1.52)	1.32 (1.23-1.41)	1.38 (1.29-1.48)
Ex-smoker vs. never	68	1.19 (1.14-1.24)	1.19 (1.14-1.25)	1.19 (1.10-1.28)	1.16 (1.08-1.25)
Total cholesterol, per 1 mmol/L increase	46	0.97 (0.95-1.00)	0.95 (0.92-0.98)	0.98 (0.95-1.01)	0.93 (0.89-0.97)
HDL, per 0.5mmol/L increase	45	0.93 (0.91-0.96)	0.95 (0.91-0.99)	0.84 (0.81-0.87)	0.87 (0.81-0.92)
Heart rate, per 10 beats/min increase	22	1.25 (1.23-1.27)	1.15 (1.13-1.17)	1.16 (1.09-1.23)	1.09 (1.06-1.12)
Creatinine, per 30µmol/L increase	62	1.21 (1.19-1.22)	1.30 (1.28-1.32)	1.24 (1.21-1.26)	1.33 (1.31-1.36)
White cell count, per 1.5 10^9/L increase	44	1.22 (1.20-1.24)	1.20 (1.18-1.21)	1.22 (1.19-1.25)	1.18 (1.15-1.21)
Haemoglobin, per 1.5 g/dL increase	47	0.66 (0.65-0.67)	0.70 (0.68-0.71)	0.72 (0.69-0.74)	0.74 (0.72-0.76)