Supplementary Material – Impact of NICE guideline change (CG181) on statin prescribing behaviour in England

Box S1 - Breakdown of predictor variables and how they were extracted from the raw electronic health record

All 'history of' variables were derived looking back in the patients' medical record from the cohort entry date for the codes from the relevant code list (below). For ethnicity the entire patients' medical record was searched as this variable cannot change.

When deriving test data we followed QRISK methods wherever possible. For test data we looked as far back as five years prior to the cohort entry date. Standard deviation of systolic blood pressure was only recorded if a patient had two or more values in the previous five years, taking the standard deviation of all recorded values. For cholesterol and HDL, we also looked forward in time up until a CVD event or censored, or five years time. Drugs at baseline were defined as at least two prescriptions, with at least one in the 28 days before the cohort entry date.

The number of prescriptions in the last year was derived from the number of distinct prescription items per day (i.e. duplicates on the same day were removed, but both counted if on different days). Number of days with medical records was the number of distinct days with a Read code in the year prior to the cohort entry date.

Severe mental illness included codes for depression, as is noted in the rapid responses section on the BMJ website: https://www.bmj.com/content/357/bmj.j2099/rapid-responses

Chronic Kidney Disease was calculated using read codes and an algorithm using test data. The algorithm uses eGFR scores and comes from this paper[1]. However many patients have creatinine recorded as opposed to eGFR. The recommended equation to convert creatinine to eGFR is the CKD-EPI[2] equation. This is recommended in the KDIGO guidelines[3], and this recent comparison[4] comparing CKD-EPI to the MDRD equation, which used to be the most commonly used equation. Therefore I extracted creatinine and eGFR scores in order to calculate presence of CKD using the above referenced algorithm.

Full algorithms to calculate BMI, SBP, SBP variability, Cholesterol/HDL ratio, smoking status and CKD are available on the GitHub page: https://github.com/alexpate30/Impact-of-Nice-guidance

Code lists for the outcome variable, cardiovascular event, were available amongst the supplementary material of the QRISK3 paper published online. For all covariates that were included in QRISK2, code lists were available from the study by Van Staa et al[5]., which compared QRISK2, ASSIGN and Framingham. I then also used the code lists available from QOF as an alternative set of code lists, given I was not sure what had been used in the QRISK3 paper.

For variables not in QRISK2, or part of QOF, code lists were not available. This was atypical anti-psychotic medication, erectile dysfunction, HIV/AIDS, migraine and systemic lupus erythematosus. For these codes were either generated through the CPRD code browser, or were available on the following websites:

http://www.phpc.cam.ac.uk/pcu/cprd_cam/codelists/[6]

https://clinicalcodes.rss.mhs.man.ac.uk/[7]

Code lists are provided on the GitHub: https://github.com/alexpate30/Impact-of-Nice-guidance

Box S2 - Breakdown of missing data and details of imputation process

Amount of missing data

The levels of missing data were as follows: cholesterol/HDL ratio [17.56% for females and 16.91% for males], SBP [1.60% and 2.26%], SBP variability [6.26% and 9.77%], Smoking [9.71% and 8.50%] and BMI [18.44% and 20.65%]. Missing data was combined with white to create a 'white or not stated' category, as is the case in QRISK3.

Imputation methods

Multiple imputation by chained equations was used to impute missing data for body mass index (BMI), systolic blood pressure (SBP) and SBP variability, cholesterol, HDL and smoking status. Missing data in Ethnicity was treated as in QRISK3, by using a 'white or not stated' category. The program used to impute the data was the R package MICE[8]. 20 imputation procedures were carried out, and 30 iterations for each one. Variables included in the imputation model were all predictor variables required to produce a risk score using QRISK3 (including interaction terms). All continuous variables were imputed using predictive mean matching, and polytomous regression for categorical variables[8]. Interactions terms were imputed empirically from the two component variables (not stochastically), and interactions terms were not used to impute their component variables.

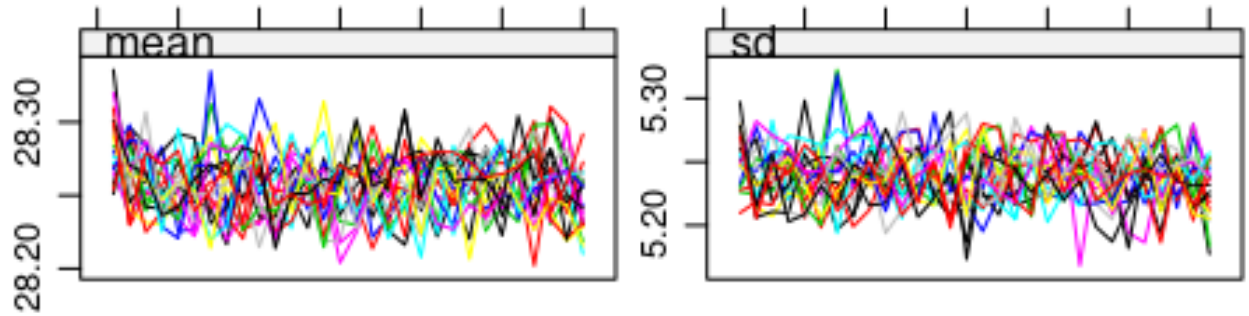
Assessment of the performance of the multiple imputation procedure

For continuous data density plots were generated to assess whether there are any systematic differences in covariates for those with missing data and those without. This also enables us to check that the distribution of imputed values is reasonable (i.e. no extreme values, or a distribution shape which clearly indicates an issue with the imputation procedure). The convergence plots assess the level of mixing in the Markov chain and whether it had reached a steady state when we drew the values. For categorical variables, the distribution of the variable from each imputation stream are presented, as well as the distribution of non-missing values.

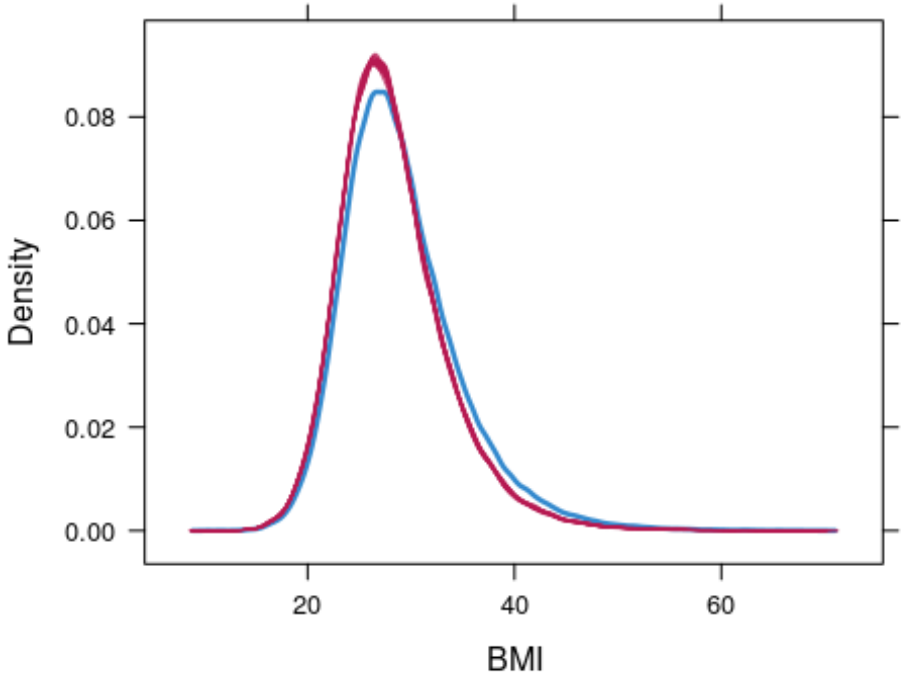
All convergence plots reached a steady state very quickly, far before the 30th iteration. All density plots seem reasonable. They are presented below.

BMI

BMI convergence plot

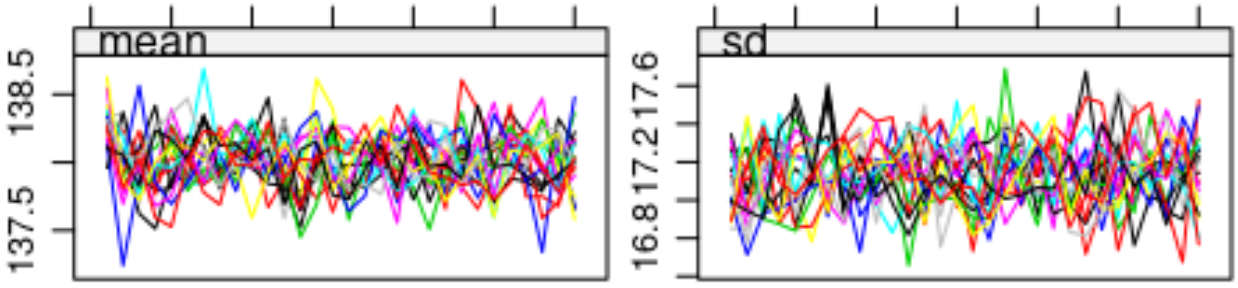


BMI density plot

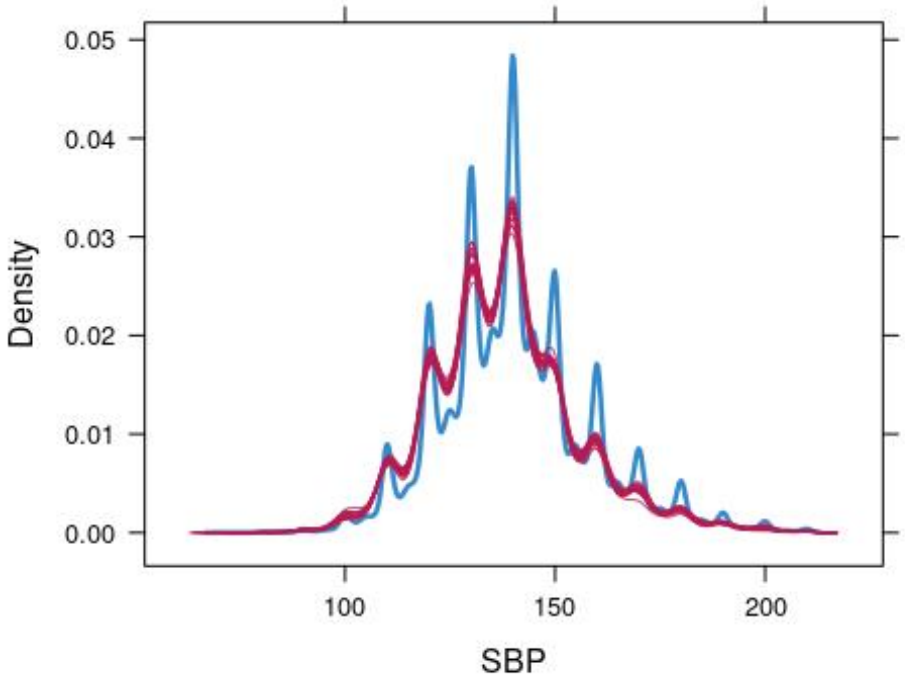


SBP

SBP convergence plot

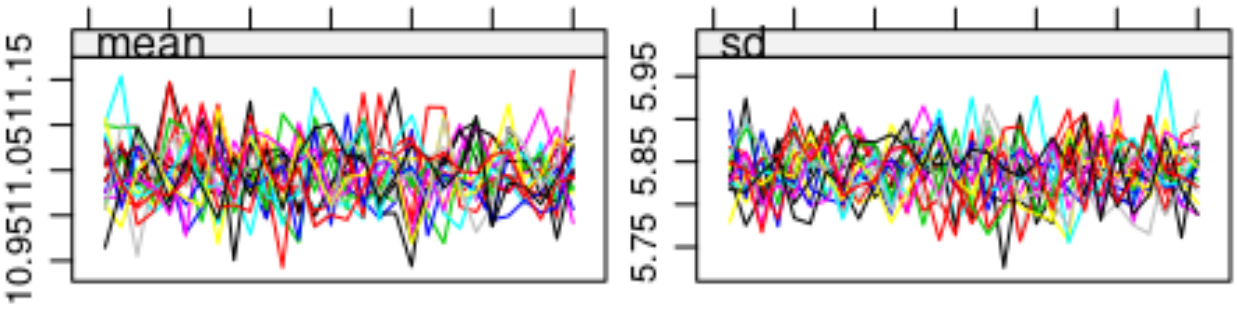


SBP density plot

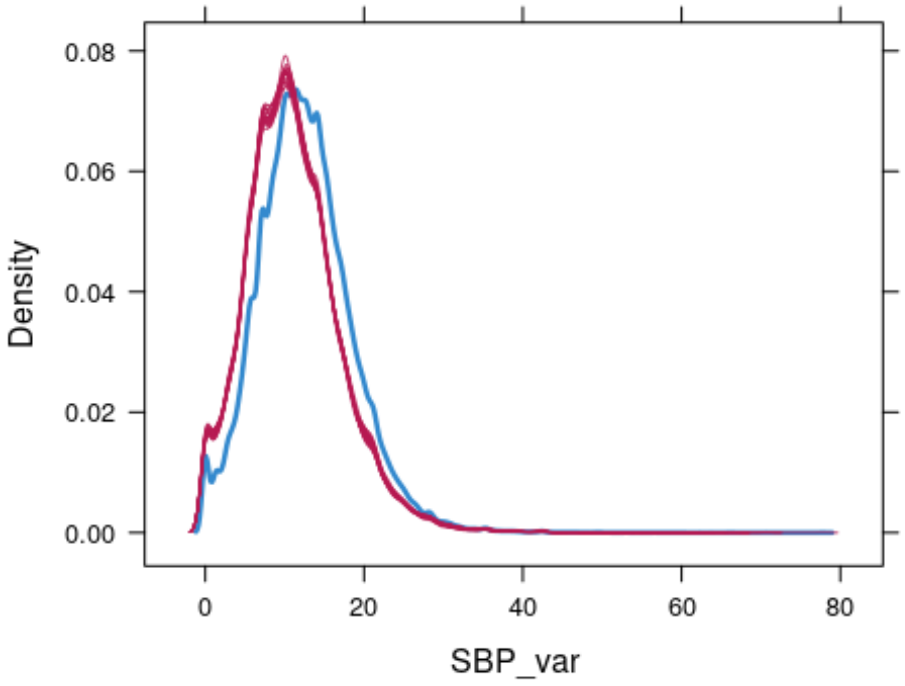


SBP variability

SBP variability convergence plot

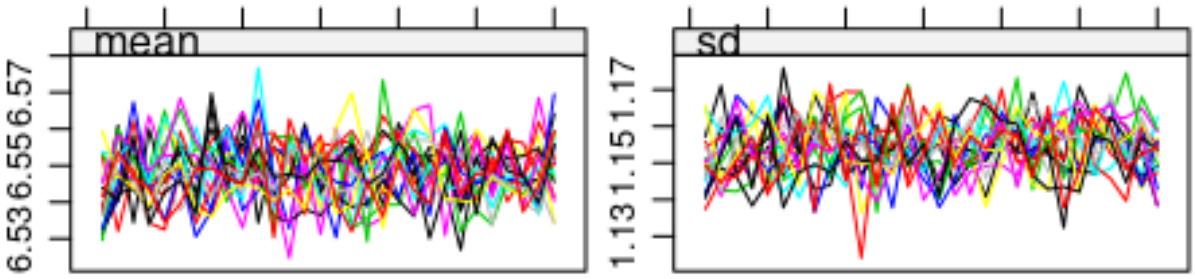


SBP variability density plot

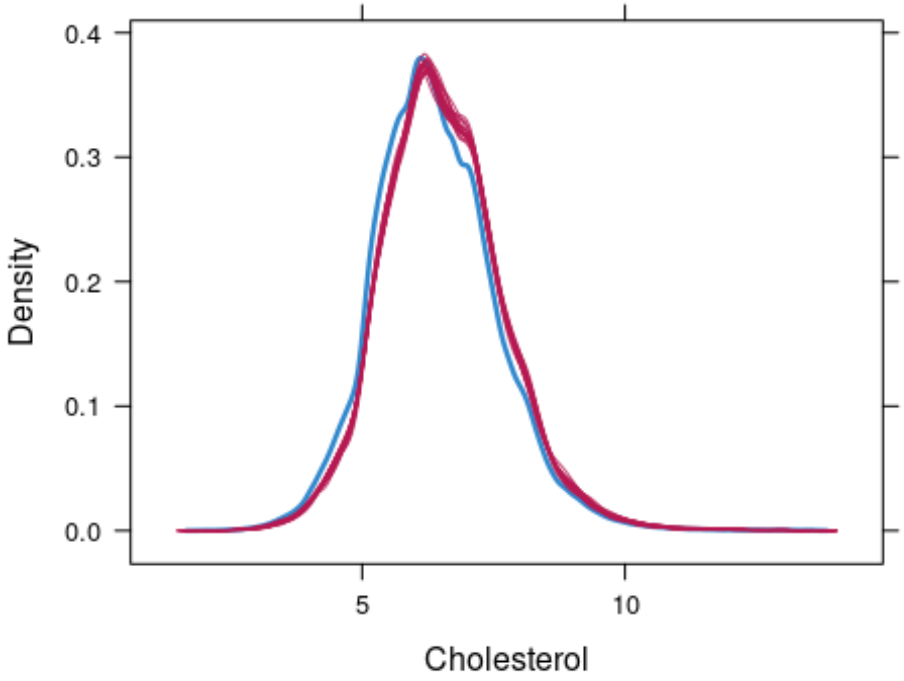


Cholesterol

Cholesterol convergence plot

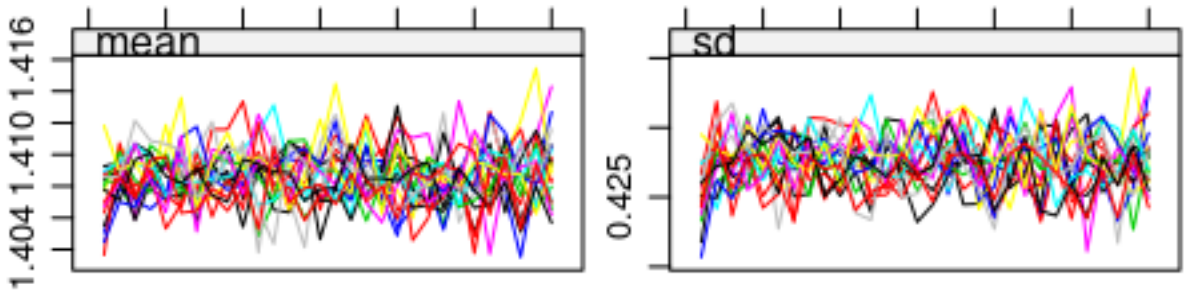


Cholesterol density plot

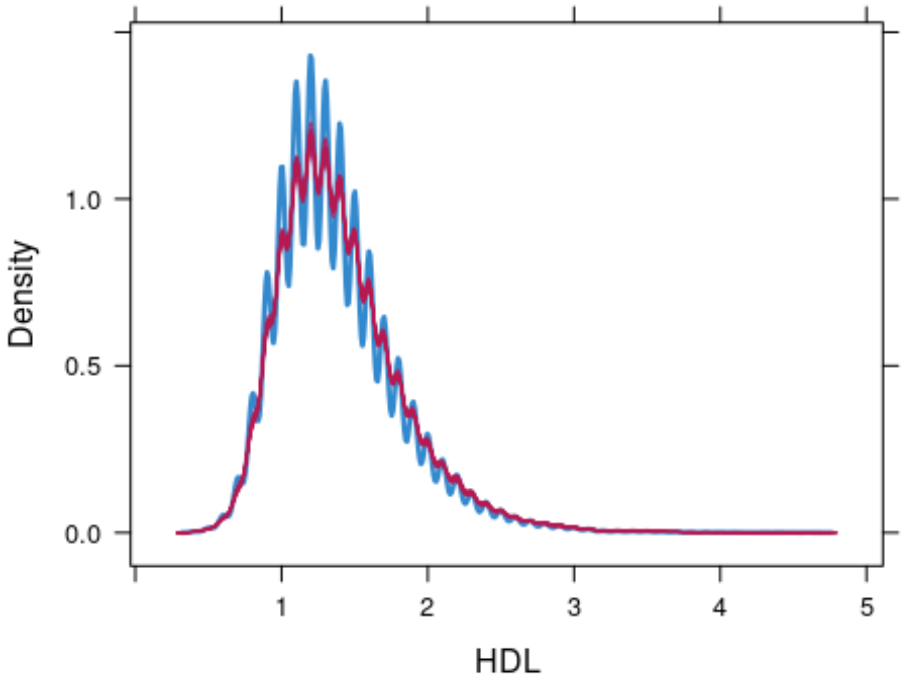


HDL

HDL convergence plot



HDL density plot



Smoking

Distribution of real data and imputed data by imputation

	Smoking status (%)				
Imputation	Never	Ex	Light	Moderate	Heavy
Real data	38.92	35.85	9.75	8.06	7.43
1	37.58	41.82	7.19	7.26	6.16
2	37.92	40.82	7.10	7.58	6.58
3	37.00	41.55	7.39	7.95	6.11
4	36.93	41.71	7.64	7.37	6.35
5	37.12	41.63	7.04	7.57	6.64
6	37.30	41.29	7.14	7.78	6.49
7	37.04	41.47	7.50	7.92	6.07
8	37.36	41.72	7.26	7.45	6.20
9	37.01	40.88	7.43	8.02	6.66
10	37.46	41.52	7.29	7.26	6.47
11	36.75	42.54	7.00	7.72	5.99
12	36.36	42.05	7.69	7.84	6.06
13	37.30	41.40	7.20	7.74	6.36
14	37.42	41.06	7.51	7.57	6.44
15	37.48	41.46	7.26	7.79	6.01
16	37.79	41.10	7.15	7.11	6.84
17	37.19	41.50	7.16	7.68	6.47
18	37.20	41.92	7.25	7.45	6.19
19	37.33	41.59	7.45	7.57	6.07
20	37.13	41.57	7.34	7.58	6.38

Table S1 - Baseline demographics of statin cohort

	Female	Male
N	166,209	185,344
Continuous variables		
Age	63.5 (11.05)	60.08 (11.08)
Systolic blood pressure	140.33 (18.35)	140.61 (17.2)
Systolic blood pressure variability	13.07 (5.8)	12.12 (5.89)
Body mass index	29.26 (6.35)	28.96 (5.05)
Cholesterol/HDL ratio	4.64 (1.42)	5.21 (1.53)
Categorical variables		
Atrial fibrillation	2.85%	3.62%
Atypical antipsychotic medication	0.86%	0.76%
Corticosteroid use	2.00%	1.23%
Chronic kidney disease stage 3/4/5	13.61%	7.17%
Diabetes (type 1)	1.32%	1.70%
Diabetes (type 2)	21.19%	22.25%
Ethnicity: Bangladesh	0.13%	0.16%
Black African	0.40%	0.40%
Black Caribbean	0.44%	0.34%
Chinese	0.13%	0.11%
Indian	0.90%	1.06%
Other	0.79%	0.83%
Other Asian	0.59%	0.64%
Pakistani	0.32%	0.39%
White	96.30%	96.08%
Family history of CVD	29.49%	23.22%
HIV	0.04%	0.13%
Treated Hypertension	48.74%	43.70%

Migraine	10.64%	4.41%
Rheumatoid Arthritis	2.10%	0.88%
Smoking: Never	46.47%	32.24%
Ex	30.60%	40.49%
Current	22.93%	27.27%
Systemic lupus erythematosus	0.23%	0.03%
Severe mental illness	15.87%	8.56%

References

1. Jameson K, Jick S, Hagberg KW, et al. Prevalence and management of chronic kidney disease in primary care patients in the UK. *Int J Clin Pract* 2014; 68: 1110–1121.
2. Levey AS, Stevens LA, Frcp C, et al. A New Equation to Estimate Glomerular Filtration Rate Andrew. *Ann Intern Med* 2009; 150: 604–612.
3. Kidney Disease: Improving Global Outcomes (KDIGO) CKD Work Group. KDIGO clinical practice guideline for the evaluation and management of chronic kidney disease. *Kidney Int Suppl* 2013; 3: 1–150.
4. Matsushita K, Woodward M, Jafar TH, et al. Comparison of Risk Prediction Using the CKD-EPI Equation and the MDRD Study Equation for Estimated Glomerular Filtration Rate. *JAMA* 2012; 307: 1941–1951.
5. Van Staa TP, Gulliford M, Ng ESW, et al. Prediction of cardiovascular risk using framingham, ASSIGN and QRISK2: How well do they predict individual rather than population risk? *PLoS One* 2014; 9: e106455.
6. Primary Care Unit CU. CPRD @ Cambridge - Code Lists, http://www.phpc.cam.ac.uk/pcu/cprd_cam/codelists/ (accessed 24 January 2018).
7. Institute of Population Health U of M. ClinicalCodes.org, https://clinicalcodes.rss.mhs.man.ac.uk/ (accessed 24 January 2018).
8. van Buuren S, Groothuis-oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. *J Stat Softw*; 45, https://www.jstatsoft.org/article/view/v045i03 (2011).