**README for programs to derive test data**

The algorithms are provided for test data only. ‘History of’ variables are derived by looking back in a patient’s medical record for specific code lists, so it is clear how these were derived. The programs themselves are not reusable, as they are specific to the server on which raw data was stored, and they are rather cumbersome. This is the same for the test data, however the test data had to undergo a cleaning procedure, where extreme values were removed and entries with different units were converted into a consistent unit. I wanted this process to be transparent, so I have provided the algorithms for deriving the test data.

**Algorithms to extract the test data BMI, SBP, SBP variability, Cholesterol/HDL ratio, smoking status and CKD are available at the GitHub page: .**

* For BMI, SBP, SBP variability, Cholesterol/HDL ratio and smoking status, the SAS programs work from the point of having already extracted all the values of a given variable (i.e. all the BMI values). These are extracted through a combination of medical codes and entities in the test and additional clinical details files. A series of data cleaning steps are then applied. The code is not directly re-usable, but the steps taken to get each variable into a consistent format should be clear from the code provided.
* For CKD, the code only relates to the portion of the variable which is identified through test data (cases also identified through medical codes). This is a two stage process. First all the GFR values are extracted from the database (SAS program: *varDeriveGFR.sas*). Some are extracted directly, and some are derived from creatinine measures using the referenced formula.(1) Then the data is exported to R, where CKD cases are identified from the GFR values (*varDeriveCKDFromGFR\_Stage45.R* or *varDeriveCKDFromGFR\_Stage345.R*) using the referenced algorithm.(2)