Plan for creating UKBioBank outcomes from csv file

Specification:

To extract incidence and prevalence of simple disease outcomes from UKBioBank data using stata via an updatable csv file.

Fields to use:

-HES fields (diag\_icd9, diag\_icd10, oper4 with date fields admidate and opdate; 40001, 40002, 41203, 41205 with death date 40000 )   
-self reporting fields (20002, 20004, 6150)   
- date fields (53, 20009, 20011 & 3627, 3894, 2966, 4056)   
- death icd fields (40001, 40002 = matching responses to HES fields)  
 - self reported medication (20003) ?

Plan:

User will input the outcome of interest and the program will return all users in the data with additional fields as follows:

Event – a binary yes/no as to whether any event matching requested values ever occurs

Prevalent/date – The earliest known event matching requested values prior to admission to biobank, using a combination of self-report of events prior to 0th survey (20002, 20004, 6150) and HES results (icd: diag\_icd9, diag\_icd10 and operation oper4) dated before 0th survey. Baseline will be taken as date of assessment of 0th survey (53). I will give age of incident if self-report (using fields 20009, 20011 & 3627, 3894, 2966, 4056 as relevant for 6150) or date of event from HES (admidate/ opdate)

Prevalent excluding self-report/date – As above, excluding the self-report data. The earliest known event prior to admission to biobank using HES results ((icd: diag\_icd9, diag\_icd10 and oper4)

Incident/date – earliest known event matching requested values after admission to biobank, using HES, death (40001, 40002) and self-reporting fields. Age at event will be reported as with prevalence. If event was not reported on an earlier survey & no age is given, then age will be set to midpoint of two survey dates. Date of death will be used for events coming from 40001 and 40002. Date of admission will be used for HES events.

Incident excluding self-report/date - as above including only events from HES or death fields (HES: (icd: diag\_icd9, diag\_icd10 and operation oper4 and 40001/40002)

Extra info:

It would be possible to include an option to save a file of all relevant events occurring, as well as reporting earliest event for each participant.

Medications – possibility of including taking medication as an indicator. Issues with assigning date to event in this case.

Important questions:

1. Admission date or episode date for date of event of HES data
2. Are there other outcomes with specialised age variables (e.g. 4012 age of DVT diagnosed) of interest

CSV format?

[code].X will match to any string starting with [code] with additional numbers

[code] will match exactly only.

Decimal points will be removed before matching.

I will be looking for events that match ANY of the codes given in ANY column.