```
/***************************************************************
*Malaria early morbidity and mortality
* 2. Data checking do file
*  First run set up
* Author: Carl Higgs    Last modified: 5 October 2015
***************************************************************/

capture log close
version 13.1
set linesize 100
set more off
cd "C:\Users\Carl\Google Drive\MPH\Projects\Malaria project\Data"
loc today = c(current_date)
log using "malariaproject_log_`today'.txt", append text



***setup for stvary  /// note: have used date variables and id options for setup for diagnostic analysis; stset for survival analysis
sort hrn doa dod

***Generate date based follow up time variables for clarity around ID times using stvary process
gen FU_Adm = doa + Adm_FU14
format %tdD_m_Y FU_Adm
gen FU_Died = doa + Died_FU14
format %tdD_m_Y FU_Died

stset FU_Adm, fail(AdmNext14) exit(FU_Adm) id(hrn) time0(doa) origin(doa)
streset, future /* includes presentations subsequent to first in the analysis*/
*Diagnostics
list hrn doa DOEnd AdmNext14 _t0 _t _d  _st if hrn<100

loc varlist obsno hrn MalCluster doa dod DOEnd Age Sex Source op ip Stay Ethnic Preg Died Species pf pv pm po MalPres TotalPres Year
Month TimeM Era MalNut Bleeding NumbDiags EthnicGr AGR2 AGR3 AGR4 aa_any cq_any cl_any dhp_any dox_any ivart_any ivq_any oralq_any
pq_any sp_any matchscript predwt PQmgkg_Sum PQdose_First PQDose_Any TreatGr_First TreatGr1_First TreatGr2_First hbmin_First hbmin_Last
 hbmin_Min pltmin wbcmin wbcmax MalariaDay Malaria_FU365 MalariaNext365 PFDay PF_FU365 PFNext365 PVDay PV_FU365 PVNext365 PMDay
PM_FU365 PMNext365 MixDay MIX_FU365 MIXNext365 AdmDay Adm_FU14 AdmNext14 SpeciesAdm DiedDay Died_FU Died_FU14 DiedNext14 Died_FU365
DiedNext365 SpeciesD MalNutD HbD HbDGr7 HbDGr5 VivaxDeath FollowUp365 Malaria_Next365 Malaria_Last14 Malaria_Last63Gr PVMIX_Include
PqDoseGr Rank_Any Rank_Pv DiedNext7 Died_FU7 constant HbAdmGr7 HbAdmGr5 sevThrom log10wbcmin log10pltmin log10PQmgkg_Sum log2wbcmin
log2pltmin log2PQmgkg_Sum q5_wbcmin q5_pltmin q5_PQmgkg_Sum dummy pregWom YearCat FU_Adm FU_Died begin_m end_Adm end_D _insmpl _st _d
_origin _t _t0

loc today = c(current_date)
loc filename = c(filename)
putexcel A1=("stvary diagnostics using `filename' on `today' " ) using diagnostics_log.xlsx, modify sheet("stvary diagnostics")
putexcel A2=("variable") B2=("constant") C2=("varying") D2=("never missing") E2=("always missing") F2=("sometimes missing") ///
    using diagnostics_log.xlsx, modify sheet("stvary diagnostics")
loc xcell = 3

foreach x of varlist `varlist' {
    stvary `x'
    putexcel A`xcell'=("`x'") B`xcell'=(r(cons)) C`xcell'=(r(varies)) D`xcell'=(r(never)) E`xcell'=(r(always)) F`xcell'=(r(miss)) ///
        using diagnostics_log.xlsx, modify sheet("stvary diagnostics")
    loc ++xcell
    }


*Post stvary diagnostics: investigate and fix where possible missing data (some may be identifiable from subsequent/prior presentations)

***dod & DOEnd ***
list hrn doa dod DOEnd AdmNext14 DiedNext14 if dod==.
list hrn doa dod DOEnd AdmNext14 DiedNext14 if inlist(hrn, 54616, 85200, 85286, 138181, 188041, 191522)
/* The missing values of dod and DOEnd were all for patients admitted on same day as presentation
(Adm_FU14 =0.4, AdmNext14==1; none died within 14 days) */

*** AGE ***
levelsof hrn if Age==., loc(hlist)
local hlist_c : subinstr local hlist " " ",", all
drop issue_Age
gen issue_Age = 0
/// Export to excel data of age and visits for hrns with missing age data; equivalent of : list hrn obsno doa Age AGR2 AGR3 AGR4 if
inlist(hrn,`hlist_c')
putexcel A1=("Diagnostics") using diagnostics_log.xlsx, modify sheet("Missing age diagnostics")
export excel hrn obsno doa Age AGR2  AGR3 AGR4 using diagnostics_log.xlsx  if inlist(hrn,`hlist_c'), cell(A2) sheetmodify
sheet("Missing age diagnostics")  firstrow(variables)  missing("*MISSING*") datestring("%tdD_m_CY")
foreach x of loc hlist {
/// Derive age based on subsequent or prior visit dates where possible
    qui: su Age if hrn==`x', meanonly
    loc n_age = r(N)

    if `n_age' == 0 {
        di "No data available from which to derive Age for missing values for HRN " `x' "."
        }
    if `n_age' > 0{
        loc minAge = r(min)
        qui: su doa if hrn==`x' & Age==`minAge' & Age!=., meanonly
        loc ageDate = r(min)
        qui misstable sum Age if hrn==`x'
        loc nm= r(N_eq_dot)
        di "hrn: " `x' "     minAge: " `minAge' "    ageDate: " %tdD_m_Y " Missing entries:" `nm'
 /// Identify details of hrn and obs with missing Age values
        list obs hrn doa Age AGR2 AGR3 AGR4 if hrn==`x'
        forval m = 1/`nm' {
            qui: su doa if hrn==`x' & Age==. , meanonly
            loc missDate = r(min)
            qui: su obsno if hrn==`x' & Age==. & doa==`missDate' , meanonly
            loc obs = r(min)
            di "obs: "  `obs' "     ageDate: " %tdD_m_Y `ageDate' "     missDate: "  %tdD_m_Y `missDate'
            if `ageDate' > `missDate' {
                replace Age = round(`minAge'-((`ageDate' - `missDate')/365.25),.1) in `obs'
                di "Age in observation `obs' = `minAge'-((`ageDate' - `missDate')/365.25) = " `minAge'-((`ageDate' - `missDate')/365.
                25)
                }
            else if `ageDate' < `missDate' {
                replace Age = round(`minAge'+((`missDate' - `ageDate')/365.25),.1) in `obs'
                di "Age in observation `obs' = `minAge'+((`missDate' - `ageDate')/365.25) = "  `minAge'+((`missDate' - `ageDate')/365.
                25)
                }
            else if `ageDate' == `missDate'{
                replace Age = `minAge' in `obs'
                di "Age in observation `obs' = " `minAge' "(same value)"
                }
replace issue_Age = 1 if Age < 0
replace issue_Age = 1 if Age ==.
replace Age = 0.01 if Age < 0 in `obs'
///Replace missing age categories based on derived age
            replace AGR2 = 1 if AGR2==. & hrn == `x' & Age < 15 in `obs'
            replace AGR2 = 0 if AGR2==. & hrn == `x' & Age >=15 & Age !=.   in `obs'
            replace AGR3 = 1 if AGR3==. & hrn == `x' & Age <5               in `obs'
            replace AGR3 = 2 if AGR3==. & hrn == `x' & Age >=5  & Age <=15  in `obs'
            replace AGR3 = 3 if AGR3==. & hrn == `x' & Age >15  & Age !=.   in `obs'
            replace AGR4 = 1 if AGR4==. & hrn == `x' & Age <1               in `obs'
            replace AGR4 = 1 if AGR4==. & hrn == `x' & Age >=1 & Age <5     in `obs'
            replace AGR4 = 1 if AGR4==. & hrn == `x' & Age >=5 & Age <15    in `obs'
            replace AGR4 = 1 if AGR4==. & hrn == `x' & Age >=15 & Age !=.   in `obs'
            sort hrn doa dod

            }
            ///Examine final changes
            di "Final changes made, if any:"
            list obs hrn doa Age AGR2 AGR3 AGR4 if hrn==`x'
        }
    }
putexcel H1=("Derivations, where possible (Age < 0 recorded as 0.01)")  using diagnostics_log.xlsx, modify sheet("Missing age
diagnostics")
export excel hrn doa Age AGR2  AGR3 AGR4 using diagnostics_log.xlsx  if inlist(hrn,`hlist_c'), cell(H2) sheetmodify  sheet("Missing
age diagnostics") firstrow(variables)  missing("*MISSING*") datestring("%tdD_m_CY")

levelsof hrn if Age==., loc(hlist)
local hlist_c : subinstr local hlist " " ",", all
list obs hrn doa Age AGR2 AGR3 AGR4 if inlist(hrn,`hlist_c')

*11/19missing ages have been derived, however for two infant patients there are negative values - it may be best to make these "0.01" but
I have not yet done this.

***Sex***
///  Examine all hrns for more than 1 response to Sex (takes approximately 40 mins - see diagnostics_log.xlsx)
/// Exports to Excel file
timer clear 1
timer on 1

sort hrn doa dod
qui: su hrn, meanonly
loc n_hrn = r(N)

putexcel A1=("HRNs with inconsistent recorded values of Sex") using diagnostics_log.xlsx, modify sheet("Inconsistent Sex")
putexcel A2=("hrn") B2=("obsno") C2=("doa") D2=("Sex")using diagnostics_log.xlsx, modify sheet("Inconsistent Sex")
loc xcell = 3
drop issue_Sex
gen issue_Sex = 0


loc varlist hrn obsno doa Sex
forval x = 1/`n_hrn' {
    qui: su Sex if hrn==`x', meanonly
    loc diffSex = r(max)-r(min)

    if `diffSex' > 0 & `diffSex' !=. {
        list `varlist' if hrn==`x'
        su hrn  if hrn==`x', meanonly
        loc n_hrn = r(N)
        export excel `varlist' using diagnostics_log.xlsx  if hrn==`x', ///
            cell(A`xcell') sheetmodify  sheet("Inconsistent Sex") missing("*MISSING*") datestring("%tdD_m_CY")
        loc xcell = `xcell'+`n_hrn'+1
        replace issue_Sex = 1 if hrn==`x'
        }
    }
codebook hrn if issue_Sex ==1

timer off 1
timer list 1

***Stay***
list hrn doa dod DOEnd AdmNext14 DiedNext14 if Stay==.
list hrn doa DOEnd Source ip Stay Adm_FU14 AdmNext14 if inlist(hrn, 54616, 85200, 85286, 138181, 188041, 191522)
/*the same patients as missing dod and DOEnd; inpatients, with 54616, 138181 &188041 having presentation through the UGD (Unit Gawat Darurat)-Emergency treatment room */

***Era***
list hrn doa Year Month Era Adm_FU14 AdmNext14  if Era==.
assert Year==2006 & Month==4 if Era==.
/*  Era is missing only for participants admitted in April 2006, presumably the month of transition into using ACT */


*** Ethnic *** (some incongruous data - participants of unique hrn listed at different times "Papuan" and "non-Papuan")
levelsof hrn if Ethnic==., loc(hlist)
/// local hlist_c : subinstr local hlist " " ",", all
/// list obs hrn doa Ethnic EthnicGr if inlist(hrn,`hlist_c')
foreach x of loc hlist {
qui: su Ethnic if hrn==`x', meanonly
loc n_Ethnic = r(N)

if `n_Ethnic' > 0 {
    list obs hrn doa Ethnic EthnicGr if hrn == `x'
    }
}

///  Examining all hrns for more than 1 response to Ethnic and EthnicGr items
sort hrn doa dod
qui: su hrn, meanonly
loc n_hrn = r(N)

putexcel A1=("HRNs with inconsistent recorded values of Ethnicity") using diagnostics_log.xlsx, modify sheet("Inconsistent ethnicity")
putexcel A2=("hrn") B2=("obsno") C2=("doa") D2=("Ethnic") E2=("EthnicGr") using diagnostics_log.xlsx, modify sheet("Inconsistent ethnicity")
loc xcell = 3
gen issue_Ethn = 0
/// The below code takes hours to run - just see diagnostics_log.xlsx for results
loc varlist hrn obsno doa Ethnic EthnicGr
forval x = 1/`n_hrn' {
    qui: su Ethnic if hrn==`x', meanonly
    loc diff1 = r(max)-r(min)
    qui: su EthnicGr if hrn==`x', meanonly
    loc diff2 = r(max)-r(min)
    if `diff1' > 0 & `diff1' !=. | `diff2' > 0 & `diff2' !=. {
        di "Ethnic difference: " `diff1' "     EthnicGr difference: " `diff2'
        list `varlist' if hrn==`x'
        qui: su hrn  if hrn==`x', meanonly
        loc n_hrn = r(N)
        export excel `varlist' using diagnostics_log.xlsx  if hrn==`x', cell(A`xcell') sheetmodify  sheet("Inconsistent ethnicity")
        missing("*MISSING*") datestring("%tdD_m_CY")
        loc xcell = `xcell'+`n_hrn'+1
        replace issue_Ethn = 1 if hrn==`x'
        }
    }
codebook hrn if issue_Ethn ==1
/// 48 unique values of hrn with issues with Ethnic or EthnicGr according to above formula

***Further examination of sex and ethnicity***
list hrn obsno doa Age Sex Ethnic EthnicGr issue_Ethn issue_Sex if issue_Ethn==1& issue_Sex ==1
/*
Problems: subject 103681 is a 2.3 year old highland female at their visit in 2013, but at their visit 7 years earlier in 2006 they are an 8.3 year old lowland boy. Similar issue with patient 103681 */

*** Compare expected  and observed ages, using age at date of first visit where age is recorded
timer clear 1
timer on 1  /* timer took 1560 seconds (26 mins) up to 7556) &  19598 seconds (5.4 hours) for remainder  */
qui: su hrn, meanonly
loc hmin = 7556/* r(min) */
loc hmax = r(max)
di "Processing hrns of " `hmin' " through " `hmax' "..."
drop predAge
drop issue_predAge
drop diffAge
gen predAge = 0
gen issue_predAge = 0
gen diffAge = 0
format %9.3f diffAge
gen long age1000 = Age*1000
```

```
forval x = `hmin'/`hmax' {
    qui: su Age if hrn==`x', meanonly
    loc n_age = r(N)
    if `n_age' > 0 {
        loc minAge = r(min)
        qui: su age1000 if hrn==`x', meanonly
        loc minAge1000 = r(min)  /* silly code to force Stata to be less precise with decimal point  */
        qui: su doa if hrn==`x' & age1000==`minAge1000'
        loc ageDate = r(min)
        levelsof obsno if hrn==`x', loc(olist)
        foreach obs of loc olist {
            qui: su doa if hrn==`x' & obsno==`obs', meanonly
            loc obsDate = r(min)
            di "hrn: " `x' "   obs: "  `obs' " minAge: " `minAge' "    ageDate: " %tdD_m_Y `ageDate' " obsDate: "  %tdD_m_Y `obsDate'
            replace predAge = round(`minAge'+((`obsDate' - `ageDate')/365.25),.1) in `obs'
            qui: su Age if obsno==`obs', meanonly
            loc r_age = r(min)
            qui: su predAge if obsno==`obs'
            loc p_age = r(min)
            replace diffAge = Age - predAge in `obs'
            loc d_age = `r_age' - `p_age'
            loc xd_age = sqrt(`d_age' * `d_age')
            if `xd_age' == 0 {
                di "For observation `obs', actual age (" `r_age' ") and predicted age (" `p_age' ") are the same," ///
                    " with difference of (" `d_age' ") year."
                }
            if `xd_age' > 0 & `xd_age'<=1 {
                di "For observation `obs', actual age (" `r_age' ") and predicted age (" `p_age' ") are similar," ///
                    " with difference of (" `d_age' ") year."
                }
            if `xd_age' >1 & `d_age' <=3 {
                di "For observation `obs', actual age (" `r_age' ") and predicted age (" `p_age' ") are different," ///
                    " with difference of (" `d_age' ") years."
                replace issue_predAge=1 in `obs'
                }
            if `xd_age' >3 & `xd_age' <.{
                di "For observation `obs', actual age (" `r_age' ") and predicted age (" `p_age' ") are very different," ///
                    " with difference of (" `d_age' ") years."
                replace issue_predAge=1 in `obs'
                }

            }
        }
    }
drop age1000
timer off 1
timer list 1
***719 hrns have predicted age over 1 year differnece
***Output inconsistencies in observed and predicted ages to excel  if difference >= 3 years

gen diffAge_pos = round(sqrt(diffAge*diffAge),0.01)
recode diffAge_pos (0/0.99 = 0 "0 to 0.99") (1/2.99 = 1 "1.00 to 2.99") (3/max = 2 "3.00 to 54.40") (.=.), gen(c_diffAge)
lab var c_diffAge "Differences between observed and expected ages (years)"
tab Year c_diffAge, matcell(diff)


*differential misclassification? prob not - just smaller sample pools
stset Adm_FU14, fail(AdmNext14)
stcox i.c_diffAge, strata(MalCluster) cluster(hrn)
stset Died_FU14, fail(DiedNext14)
stcox i.c_diffAge, strata(MalCluster) cluster(hrn)

*** Possible problem with Species - mixed designation (as described in RSMM on p 4)
codebook obs if pf==1 & pv==1 & Species!=5
/*There are 1017 cases which appear to have both pf and pv infection recorded, but are not mixed - should they be?
    Answer: the Species variable is initial infection; possible to have subsequently acquired (but unlikely in 14 day period).
 */
 *** Pregnant males
 list hrn obsno Sex Preg if Sex==1 & Preg==1
 list hrn obsno Sex sexPreg Preg if inlist(hrn, 88007, 108005)

 replace Preg = 0 in 57271
replace Preg = 0 in 79337
replace sexPreg = 1 in 79337
replace sexPreg = 1 in 57271
list hrn Sex sexPreg Preg if inlist(hrn, 88007, 108005)
```