```

version 13.1

    log using /*redacted*/ , append smcl name(DataReviewNovember2017)
    *   Saurabh Chavan

    *   This code will examine the prepared data files for /*redacted*/ for inconsistencies and incompatibilities with the /*redacted*/ data requirements

    * Tables are examined in alphabetical order

    clear
    capture

    *** /*redacted*/ ***

    * We are not submitting /*redacted*/ table this time
    * It could be a consideration for the March 2018 or July 2018 upload after a review of /*redacted*/ and /*redacted*/ data

    /*use *redacted*

        sort sitepatientid
        tab datagroup,m

    clear
    capture
*/

    *** /*redacted*/ ***

    * this table can be used to delete erroneous records that were uploaded earlier

    use /*redacted*/,clear
        tab recordtype,m


    clear
    capture

    *** /*redacted*/ ***

    use /*redacted*/
        sort sitepatientid

        *   gender
        tab presentsex birthsex,m row col
        tab transgender,m
        table presentsex birthsex, by(transgendered)
        gen checkgender=1 if presentsex==birthsex & transgendered=="Yes" & presentsex!=""
        /* checking for inconsistent gender */
        tab checkgender,m
        drop checkgender

        *   death
        tab deathdate if deathdate<(date("04-01-2000","MDY")) & deathdate>(date("10-31-2017","MDY")),m
        /* to check implausible deathdates if before cohort start date or after database cut date */
        count if deathdate!=.
        *   /*redacted*/ deceased patients
        tab deathdatesource deathdateprecision ,m
        gen deathyear=yofd(deathdate)
        tab deathdatesource,m

        tab2xl deathyear using /*redacted*/, col(1) row(1) replace
        histogram deathyear
        tab deathyear, plot

        *   birthyear/age
        tab2xl birthyear using /*redacted*/, col(1) row(1) replace
        tab birthyear,m
        gen age=2017-birthyear if deathdate==.
        replace age=deathyear-birthyear if deathdate!=.
        tab2xl age using /*redacted*/, col(1) row(1) replace
        summarize age, detail
        histogram age, normal

        *   race/ethnicity
        tab race hispanic,m row col
        list sitepatientid if race=="" & hispanic==""

    preserve
        keep if race=="" & hispanic==""
        /* missing race/ethnicity */
        destring sitepatientid, replace force
        merge 1:1 sitepatientid using /*redacted*/
        keep if _merge==3
        keep mrn sitepatientid pat_lname pat_fname lastname firstname dob ssn newpatient
        export excel using /*redacted*/, sheet("missingracehisp") firstrow(var)  replace
    restore

        tab birthcountry,m
        drop age deathyear

    clear

    ***  /*redacted*/ ***
    clear

    use /*redacted*/,clear

        sort sitepatientid siterecordid

        tab datasource,m
        tab diagnosisdateprecision,m

        codebook sitepatientid
    * /*redacted*/ patients have any diagnosis

        duplicates tag sitepatientid diagnosisname diagnosisdate, gen(dupdx)
        /* duplicate diagnoses */
        tab dupdx,m

    preserve
        keep if dupdx>0
        gsort -dupdx sitepatientid diagnosisdate diagnosisname
    capture: export excel using /*redacted*/, sheet("duplicatedx") firstrow(var) sheetmodify
    restore

    drop dupdx
    preserve

```

```
        gen dxproblemdate="checkfulldate" if diagnosisdateprecision=="unknown" & diagnosisdate!=date("01/01/1900","MDY")
    * this will mark all non "01/01/1900" dates that have unknown precision (which it shouldn't be)

        replace dxproblemdate="checkmonth" if diagnosisdateprecision=="year" & month(diagnosisdate)!=1
    * this will mark something like 04/01/2011 or 04/04/2011 since year precision is only 01/01/someyear and nothing else

        replace dxproblemdate="checkday" if diagnosisdateprecision=="year" & day(diagnosisdate)!=1
    * this will mark something like 01/04/2011 or 04/04/2011 since year precision is only 01/01/someyear and nothing else

        replace dxproblemdate="checkdaymonth" if diagnosisdateprecision=="year" & day(diagnosisdate)!=1 & month(diagnosisdate)!=1
    * this will remark 04/04/2011 from above since month precision is always somemonth/01/someyear and nothing else

        replace dxproblemdate="checkday" if diagnosisdateprecision=="month" & day(diagnosisdate)!=1
        keep if dxproblemdate!=""

    capture: export excel using /*redacted*/, sheet("dxproblemdate") firstrow(var) sheetmodify
    save /*redacted*/, replace
    restore,preserve

        gen dxdateCR="01jananyyear" if diagnosisdateprecision=="year" & day(diagnosisdate)==1 & month(diagnosisdate)==1
    * this will mark something like jan/01/2000 to check if it is not underprecise (could be month or day instead)

        replace dxdateCR="01anymonanyyear" if diagnosisdateprecision=="month" & inlist(month(diagnosisdate),1,2,3,4,5,6,7,8,9,10,11,12) & day(diagnosisdate)==1
    * this will mark something like sep/01/1990 to check if it is not underprecise (could be day instead)

        replace dxdateCR="01anymonanyyear" if diagnosisdateprecision=="day" & inlist(month(diagnosisdate),1,2,3,4,5,6,7,8,9,10,11,12) & day(diagnosisdate)==1
    * this will mark something like sep/01/1990 to check if it is not overprecise (could be month instead)

        keep if dxdateCR!=""
        sort sitepatientid diagnosisdate diagnosisname
    capture: export excel using /*redacted*/, sheet("dxdatechartreview") firstrow(var) sheetmodify

    restore
    preserve

        merge m:1 sitepatientid using /*redacted*/
        drop if _merge==2
        drop _merge

        gen dxafterlastvisit=1 if encounterdate<diagnosisdate & diagnosisdateprecision=="day" & yofd(encounterdate)<2016

    * there should not be any diagnoses after the last visit date if the visit is from 2016 or before
    * is it possible to have a diagnosis date months after the last recorded visit?
    * it is possible in 2017 if the diagnosis and visits table were not cut for the same
    * it is possible in 2017 if the diagnosis and visits table were not cut for the same last through date

        tab dxafterlastvisit,m
        sort dxafterlastvisit sitepatientid diagnosisdate diagnosisname
        keep if dxafterlastvisit==1
        keep siterecordid sitepatientid diagnosisname diagnosisdate diagnosisdateprecision datasource historical encounterdate encountertype department encounterlocation dxafterlastvisit
        sort sitepatientid diagnosisdate diagnosisname encounterdate

    capture: export excel using /*redacted*/,  sheet("dxafterlastvisit") firstrow(var) sheetmodify
    restore
    preserve

    * there should be no diagnoses after deathdate and there aren't any

        merge m:1 sitepatientid using /*redacted*/
        keep if _merge==3
        drop _merge
         gen dataafterdeathdate=1 if (deathdate<diagnosisdate) & deathdate!=. & diagnosisdate!=. & diagnosisdateprecision=="day" & diagnosisdate!=date("01/01/1900","MDY")
         sort dataafterdeathdate
         tab dataafterdeathdate,m
    capture: export excel using /*redacted*/, firstrow(var) sheet("dxafterdeath") sheetmodify
    restore

        gen dxyear=yofd(diagnosisdate) if yofd(diagnosisdate)!=1900
        histogram dxyear
        tab dxyear, plot

        tab2xl dxyear using /*redacted*/, row(1) col(1)
        tab historical,m


    * there are multiple diagnoses in duplicate and triplicate and up to 9 on the same day - simply because the person has more than one visit on that particular day.
    * /*redacted*/ wants all diagnosis dates but are they ok with more than two-three rows per day?
    * Confirm with /*redacted*/?


    *** /*redacted*/ ***

    clear
    use /*redacted*/
        codebook sitepatientid

    preserve
        duplicates drop sitepatientid testdate, force
        sort sitepatientid testdate
        by sitepatientid: gen count=_N
        tab count,m
    restore

        tab mutation, sort
    * capture top ten mutations


        tab mutation if mutation=="NULL"
    preserve
    preserve
        sort testdate
        gen rank=_n
        list if rank==1
    restore,preserve
        gsort -testdate
        gen rank=_n
        list if rank==1
    restore

        bysort sitepatientid testdate: gen testcount=_N
        bysort sitepatientid testdate: gen testrank=_n
        by sitepatientid: gen testyear=yofd(testdate) if testcount==testrank

        tab2xl testyear using /*redacted*/, col(1) row(1) replace
        tab testyear, plot

    * first test in 2001-06-27
    * last test in 2015-04-27

    * none in 2016 - 2017?
    * what about those that joined the cohort after April 2015?
    * what about tests before June 2001? /*redacted*/ has said we let go of these
    * prioritise for /*redacted*/; waiting to hear from data team


    *** /*redacted*/ ***

    clear
    use /*redacted*/

        codebook sitepatientid
        codebook sitepatientid if zipcode=="ZZZZZ"

        codebook sitepatientid if (real(zipcode)>94102 & real(zipcode)<94188) & real(zipcode)!=.

    preserve
        keep if (real(zipcode)>94102 & real(zipcode)<94188) & real(zipcode)!=.
        /* local zipcodes */
        gen zip=real(zipcode)
        duplicates drop sitepatientid zip, force

        tab2xl zip using /*redacted*/, row(1) col(1) replace
    restore

    ***  /*redacted*/ ***

    clear
    use /*redacted*/
        * gen double adm=clock(admitdate,"MDYhms")
        * gen double dsc=clock(dischargedate,"MDYhms")
        * format adm %tc
        * format dsc %tc
        codebook sitepatientid

    preserve
        tab admitdateprecision,m
        tab dischargedateprecision,m

        sort sitepatientid admitdate
        gen checkdschdate=1 if dischargedate<admitdate
        tab checkdschdate,m
        drop checkdschdate
        gen longstay=">6mon" if ((dischargedate-admitdate)/30)>6
        tab longstay,m
        codebook sitepatientid if longstay!=""
    * /*redacted*/ patients had /*redacted*/ admissions with stays longer than 6 months
        gen hospstay=(dischargedate-admitdate)
        gsort -hospstay
        gsort sitepatientid  -hospstay   longstay admitdate

        sort admitdate
        gen rank=_n
        egen firstadm=min(rank) if rank==1
        list sitepatientid siterecordid admitdate admitdateprecision dischargedate dischargedateprecision if firstadm==1
        gsort -admitdate
        replace rank=_n
        egen lastadm=min(rank) if rank==1
        list sitepatientid siterecordid admitdate admitdateprecision dischargedate dischargedateprecision if lastadm==1

        sort dischargedate
        replace rank=_n
        egen firstdsc=min(rank) if rank==1
        list sitepatientid siterecordid admitdate admitdateprecision dischargedate dischargedateprecision if firstdsc==1
        gsort -dischargedate
        replace rank=_n
        egen lastdsc=min(rank) if rank==1
        list sitepatientid siterecordid admitdate admitdateprecision dischargedate dischargedateprecision if lastdsc==1

        drop rank
        sort sitepatientid admitdate
        by sitepatientid: gen rank=_n
        summarize rank, detail
        summarize hospstay, detail

        gen admityear=yofd(admitdate)
        tab admityear,m plot
        tab2xl admityear using /*redacted*/, row(1) col(1) replace
        bysort admityear: gen count=_N
        duplicates drop admityear,force
        keep admityear count

    * histogram of admissions by year

        graph bar (sum) count, over(admityear)

    restore

    * no observations from 2000,2001?? /*redacted*/ said this is all right. We let it go


    ***  /*redacted*/ ***

    * why are all stop dates unknown???
    * we are submitting only one insurance type for each patient, that too the most recent
    * /*redacted*/ allows multiple types in a chronological order, while requesting the insurance at the time of the initial visit to be sent if sites are only sending one record per patient

    clear
    use /*redacted*/
        * gen int insstadate=dofc(clock(insurancestartdate,"MDYhms"))
        * format insstadate %td
        * gen int insstodate=dofc(clock(insurancestopdate,"MDYhms"))
        * format insstodate %td
        sort sitepatientid insurancestartdate insurancetype

        tab insurance,m
        tab insurancestartdateprecision insurancestopdateprecision,m
        table insurancestartdateprecision insurancestopdateprecision,m by(insurancetype)

    clear

    ***  /*redacted*/ ***

    use /*redacted*/


        * gen double resdatetime=clock(resultdate,"MDYhms")
        * format resdatetime %tc
        * gen resdate=dofc(clock(resultdate,"MDYhms"))
        * format resdate %td


    * to check the first and last lab dates*/
        sort resultdate
        gen rank=_n
        egen firstlab=min(rank) if resultdate!=date("01/01/1900","MDY")
        egen lastlab=max(rank)
        list if firstlab==rank | lastlab==rank
        drop firstlab lastlab rank

    preserve
        gen labproblemdate="checkfulldate" if resultdateprecision=="unknown" & resultdate
!=date("01/01/1900","MDY")
    * this will mark all non "01/01/1900" dates that have unknown precision (which it
shouldn't be)

        replace labproblemdate="checkmonth" if resultdateprecision=="year" & month(
resultdate)!=1
    * this will mark something like 04/01/2011 or 04/04/2011 since year precision is only
01/01/someyear and nothing else

        replace labproblemdate="checkday" if resultdateprecision=="year" & day(resultdate
)!=1
    * this will mark something like 01/04/2011 or 04/04/2011 since year precision is only
01/01/someyear and nothing else

        replace labproblemdate="checkdaymonth" if resultdateprecision=="year" & day(
resultdate)!=1 & month(resultdate)!=1
    * this will remark 04/04/2011 from above since month precision is always
somemonth/01/someyear and nothing else

        replace labproblemdate="checkday" if resultdateprecision=="month" & day(resultdate
)!=1
        keep if labproblemdate!=""
        tab labproblemdate,m
        capture: export excel using /*redacted*/, sheet("labdatechartreview1") firstrow(
var)  sheetmodify
    restore,preserve
        gen resdateCR="01jananyyear" if resultdateprecision=="year" & day(resultdate)==1 &
 month(resultdate)==1
```

```
    * this will mark something like jan/01/2000 to check if it is not underprecise (could be month or day instead)

        replace resdateCR="01anymonanyyear" if resultdateprecision=="month" & inlist(month(resultdate),1,2,3,4,5,6,7,8,9,10,11,12) & day(resultdate)==1
    * this will mark something like sep/01/1990 to check if it is not underprecise (could be day instead)

        replace resdateCR="01anymonanyyear" if resultdateprecision=="day" & inlist(month(resultdate),1,2,3,4,5,6,7,8,9,10,11,12) & day(resultdate)==1
    * this will mark something like sep/01/1990 to check if it is not overprecise (could be month instead)

        keep if resdateCR!=""
        tab resdateCR,m
        capture: export excel using /*redacted*/, sheet("labdatechartreview2") firstrow(var)  sheetmodify
    restore,preserve

    * these issues above were not observed

        encode testname, gen(name)
        labelbook name
        tab name,m
        drop name
        codebook testname
    * trailing blanks in many responses*/
        tab result if testname=="Height",m
        tab result if testname=="Weight",m
        tab result if testname=="CD4 cell absolute",m

        tab units,m
    * are non uniform units ok?
        egen minmax=concat(normalmin normalmax unit testname),punct(" ")
        table minmax,m
        drop minmax
        tab historical,m
    restore,preserve
        merge m:1 sitepatientid using /*redacted*/
        gen dataafterdeathdate=1 if (deathdate<dofc(resultdate)) & deathdate!=. & resultdate!=. & resultdateprecision=="day" & dofc(resultdate)!=date("01/01/1900","MDY")
        tab dataafterdeathdate _merge,m
    restore,preserve
    * to identify potential misclassification of the data source
    * if precision is time then how come data source is "Source unknown"?

        tab datasource resultdateprecision,m
        codebook resultdate if datasource=="Source unknown" & resultdateprecision=="time"
        gen checksource=1 if datasource=="Source unknown" & resultdateprecision=="time"
        tab checksource,m
        keep if checksource!=.
        keep sitepatientid result resultdate testname
        capture: export excel using /*redacted*/, sheet("labsourceCR") firstrow(var) sheetmodify
    restore,preserve
        codebook sitepatientid if strmatch(testname,"*HIV*")==1
        codebook sitepatientid
    * /*redacted*/ patients have any HIV test - /*redacted*/ have any tests while /*redacted*/ patients have no tests at all - possible that they were new at the time of dataset cutting and have had tests elsewhere and no test at /*redacted*/ yet and no tests at all (/*redacted*/)
    * what to do if patients qualify for /*redacted*/ but have not had HIV test at /*redacted*/ for months?
```

```
/*redacted*/ for months?

        codebook sitepatientid if strmatch(testname,"*CD4 cell absolute*")==1
        codebook sitepatientid if strmatch(testname,"*HIV-1 RNA*")==1 | strmatch(testname,"*HIV-1 Viral*")==1
    * /*redacted*/ have at least one CD4 count
        sort sitepatientid resultdate
    restore

    preserve
        keep if strmatch(testname,"*CD4 cell absolute*")==1
        destring result, replace force
        sort sitepatientid resultdate
        by sitepatientid: gen rank=_n
        keep if rank==1
        summarize result, detail
    restore,preserve
        keep if strmatch(testname,"*HIV-1 RNA*")==1 | strmatch(testname,"*HIV-1 Viral*")==1
        replace result = regexs(0) if regexm(result, "[0-9]*$") /*to remove < > = from the results*/
        destring result, replace force
        sort sitepatientid resultdate
        by sitepatientid: gen rank=_n
        keep if rank==1
        summarize result, detail
    restore


    clear

    *** /*redacted*/ ***

    clear

    use /*redacted*/, clear
    preserve
        duplicates tag sitepatientid startdate medicationname, gen (duplicatemeds)
        tab duplicatemeds,m
        sort sitepatientid startdate medicationname
        sort medicationname
        merge m:1 medicationname using /*redacted*/.dta
        drop if _merge==2
        destring sitepatientid, replace force
        sort sitepatientid startdate medicationname
        by sitepatientid startdate: gen artcount=1 if code=="ART"
        by sitepatientid startdate: egen maxart=sum(artcount) if startdate!=date("01/01/1900","MDY") | startdate!=enddate
        tab maxart,m /* maximum number of ARTs prescribed */
        keep maxart sitepatientid medicationname startdate startdateprecision enddate enddateprecision
        gsort -maxart sitepatientid startdate medicationname
        keep if maxart!=.
    capture: export excel using /*redacted*/, sheet("maxART") firstrow(var)  sheetmodify
    restore
    preserve
        gen checkenddate=0
        replace checkenddate=1 if enddate < startdate & enddate!=date("01/01/1900","MDY") & enddateprecision=="day"
        tab checkenddate,m /*this will tag erroneous end dates that are earlier than startdates*/
         foreach i in bcheckenddate {
        drop if `i'==0
        sort sitepatientid startdate
        sort sitepatientid startdate
        save /*redacted*/, replace
            capture: export excel using /*redacted*/, sheet("checkenddatemeds") firstrow(var)  sheetmodify
        restore,preserve
}
    restore
    preserve
    * to spot bad start dates*/
        gen sproblemdate="checkfulldate" if startdateprecision=="unknown" & startdate!=date("01/01/1900","MDY")
    * this will mark all non "01/01/1900" dates that have unknown precision (which it shouldn't be)

        replace sproblemdate="checkmonth" if startdateprecision=="year" & month(startdate)!=1
    * this will mark something like 04/01/2011 or 04/04/2011 since year precision is only 01/01/someyear and nothing else

        replace sproblemdate="checkday" if startdateprecision=="year" & day(startdate)!=1
    * this will mark something like 01/04/2011 or 04/04/2011 since year precision is only 01/01/someyear and nothing else

        replace sproblemdate="checkdaymonth" if startdateprecision=="year" & day(startdate)!=1 & month(startdate)!=1
    * this will remark 04/04/2011 from above since month precision is always somemonth/01/someyear and nothing else

        replace sproblemdate="checkday" if startdateprecision=="month" & day(startdate)!=1
        tab sproblemdate,m
        sort startdate sitepatientid
          foreach name in sproblemdate  {
          drop if sproblemdate==""
          sort sproblemdate startdate sitepatientid
          save /*redacted*/, replace
          capture: export excel using /*redacted*/, sheet("problemprecisionstartmeds") firstrow(var)  sheetmodify
    restore, preserve
}
    restore
    preserve
        gen sdateCR="01jananyyear" if startdateprecision=="year" & day(startdate)==1 & month(startdate)==1
        * this will mark something like jan/01/2000 to check if it is not underprecise (could be month or day instead)

        replace sdateCR="01anymonanyyear" if startdateprecision=="month" & inlist(month(startdate),1,2,3,4,5,6,7,8,9,10,11,12) & day(startdate)==1
        * this will mark something like sep/01/1990 to check if it is not underprecise (could be day instead)

        replace sdateCR="01anymonanyyear" if startdateprecision=="day" & inlist(month(startdate),1,2,3,4,5,6,7,8,9,10,11,12) & day(startdate)==1
        * this will mark something like sep/01/1990 to check if it is not overprecise (could be month instead)

        tab sdateCR,m
        sort startdate sitepatientid
          foreach name in sdateCR  {
                  drop if sdateCR==""
                  sort sdateCR startdate sitepatientid
     save /*redacted*/, replace
    capture: export excel using /*redacted*/, sheet("medstartdateCR") firstrow(var)
```

```
sheetmodify
    restore, preserve
}
    restore
    preserve
    * to spot bad end dates*/
        gen eproblemdate="checkfulldate" if enddateprecision=="unknown" & enddate!=date("01/01/1900","MDY")
    * this will mark all non "01/01/1900" dates that have unknown precision (which it shouldn't be)

        replace eproblemdate="checkmonth" if enddateprecision=="year" & month(enddate)!=1
    * this will mark something like 04/01/2011 or 04/04/2011 since year precision is only 01/01/someyear and nothing else

        replace eproblemdate="checkday" if enddateprecision=="year" & day(enddate)!=1
    * this will mark something like 01/04/2011 or 04/04/2011 since year precision is only 01/01/someyear and nothing else

        replace eproblemdate="checkdaymonth" if enddateprecision=="year" & day(enddate)!=1 & month(enddate)!=1
    * this will remark 04/04/2011 from above since month precision is always somemonth/01/someyear and nothing else
        replace eproblemdate="checkday" if enddateprecision=="month" & day(enddate)!=1
        tab eproblemdate,m
        sort startdate sitepatientid
          foreach name in eproblemdate  {
                  drop if eproblemdate==""
                  sort eproblemdate startdate sitepatientid
     save /*redacted*/, replace
    capture: export excel using /*redacted*/, sheet("problemprecisionendmeds") firstrow(var)  sheetmodify
          restore, preserve
}
    restore
    preserve
        gen edateCR="01jananyyear" if enddateprecision=="year" & day(enddate)==1 & month(enddate)==1
    * this will mark something like jan/01/2000 to check if it is not underprecise (could be month or day instead)

        replace edateCR="01anymonanyyear" if enddateprecision=="month" & inlist(month(enddate),1,2,3,4,5,6,7,8,9,10,11,12) & day(enddate)==1
    * this will mark something like sep/01/1990 to check if it is not underprecise (could be day instead)

        replace edateCR="01anymonanyyear" if enddateprecision=="day" & inlist(month(enddate),1,2,3,4,5,6,7,8,9,10,11,12) & day(enddate)==1
    * this will mark something like sep/01/1990 to check if it is not overprecise (could be month instead)

    tab edateCR,m
    sort startdate sitepatientid
      foreach name in edateCR  {
      drop if edateCR==""
      sort edateCR startdate sitepatientid
    save /*redacted*/, replace
    capture: export excel using /*redacted*/, sheet("medenddateCR") firstrow(var) sheetmodify
    restore, preserve
}
    restore

    preserve
    * to identify startdates==enddates
        gen samedate="NO"
        replace samedate="YES" if startdate==enddate & startdateprecision=="day" & enddateprecision=="day"
        replace samedate="OK ongoing" if samedate=="YES" & enddatetype=="Ongoing"
        replace samedate="OK statmed" if samedate=="YES" & sig=="Stat"

        table enddatetype samedate, by(datasource)
        sort samedate startdate startdateprecision
    * this will give an idea of how these three interrelate and should raise appropriate suspicions for certain combinations
    restore

    preserve
        duplicates drop medicationname,force
        sort medicationname
        gen rank=_n
        keep medicationname rank
    save /*redacted*/,replace
    clear
    import excel using /*redacted*/, sheet("standardCodes_Medication") firstrow
        sort code
        rename code medicationname
        merge medicationname using medicationname
        tab _merge
        keep category medicationname _merge
        gen medstatus="in CNICS" if _merge==1
        replace medstatus="in Upload" if _merge==2
        replace medstatus="in both" if _merge==3
        drop _merge
        sort medstatus category medicationname
    save /*redacted*/,replace

    restore
    *   to indirectly confirm whether or not all CNICS required medications are covered

        tab form route,m
    *   to identify implausible combinations like injectable pills*/
    preserve
        gen checkformroute=0
        replace checkformroute=1 if route=="IM" & inlist(form,"injection","injectable solution","solution")==0
        replace checkformroute=1 if route=="IV" & inlist(form,"injection","injectable solution","solution")==0
        replace checkformroute=1 if route=="PO" & inlist(form,"pill","troches","syrup","tablet","capsule")==0
        replace checkformroute=1 if route=="PR" & inlist(form,"suppository")==0
        replace checkformroute=1 if route=="PV" & inlist(form,"cream","suppository")==0
        replace checkformroute=1 if route=="SL" & inlist(form,"pill")==0
        replace checkformroute=1 if route=="inhalation" & inlist(form,"puff","inhaler","liquid")==0
        replace checkformroute=1 if route=="intranasal" & inlist(form,"puff")==0
        replace checkformroute=1 if inlist(route,"intraocular (right)","intraocular","intraocular (left)")==1 & inlist(form,"solution")==0
        replace checkformroute=1 if route=="sub-Q" & inlist(form,"injection","injectable solution")==0
        replace checkformroute=1 if route=="topical" & inlist(form,"cream","gel","liquid","lotion","ointment","patch","solution")==0

    *   since route and form both may be missclassified, one needs to check what the
```

```
medication is in order to make the correct determination for both
        tab checkformroute,m
        tab strength,m
        tab units,m
        tab sig,m
        tab enddatetype,m
        tab enddateprecision,m
        tab datasource,m
        tab stopreason,m
        tab historical,m

    * to see the first and the last medication prescribed
        sort startdate startdateprecision
        gen rank=_n
        egen firstmed=min(rank) if startdate!=date("01/01/1900","MDY")
        list sitepatientid medicationname startdate startdateprecision enddate enddateprecision enddatetype datasource if rank==firstmed

        gsort -startdate
        replace rank=_n
        egen lastmed=min(rank) if rank==1
        list sitepatientid medicationname startdate startdateprecision enddate enddateprecision enddatetype datasource if lastmed==1
    restore
    clear


    ***  /*redacted*/ ***
    * we have not submitted this table traditionally. we submit the medication table

    ***  /*redacted*/ ***
    use /*redacted*/
        tab CODcodelabel,m
        tab type,m
        tab source,m

    ***  pro ***

    * as this table is generated through a separate and independent automated process, there isn't much scope for inconsistencies to creep in as such
    clear
    use /*redacted*/
        sort sitepatientid siterecordid
        tab projectid,m
        tab sessionid,m
        tab questionid,m
        tab sequence,m
        tab state,m
        tab value,m
    clear


    ***  procedure ***

    use /*redacted*/,clear
        tab siteprocedure,m
        codebook sitepatientid

    clear

    ***  /*redacted*/ ***
    * two possible inconsistencies in this table are duplicate risks and incorrectly coded risks

    clear
    use /*redacted*/,clear
    preserve
        sort sitepatientid siterecordid
        encode risk, gen(risktype)
        labelbook risktype
        tab risk,m

        * check if properly coded as /*redacted*/
        duplicates tag sitepatientid risk, gen(duprisk)
        tab duprisk,m

        bysort sitepatientid: gen riskcount=_n
        /* maximum number of discrete risks */
        bysort sitepatientid: egen maxrisk=max(riskcount)

        count if riskcount==maxrisk & riskcount==1
        count if riskcount==maxrisk & riskcount==2
        count if riskcount==maxrisk & riskcount==3
        count if riskcount==maxrisk & riskcount==4

        codebook sitepatientid
    * /*redacted*/ of /*redacted*/ have risks recorded - please see /*redacted*/
    restore

    clear

    ***  specimenTracking   ***
    clear

    use /*redacted*/,clear
        gen datecol=date(datecollected,"MDY")
        drop datecollected
        rename datecol datecollected
        gen dateproc=date(dateprocessed,"MDY")
        drop dateprocessed
        rename dateproc dateprocessed
        format datecollected dateprocessed %tdCY-N-D
    preserve
        sort datecollected
        gen rank=_n
        list if rank==1
        gsort -datecollected
        replace rank=_n
        list if rank==1
        drop rank
        gen colyear=year(datecollected)
        gen procyear=year(dateprocessed)
        tab colyear,m
        tab procyear,m
        tab2xl colyear using /*redacted*/, col(1) row(1) replace
        tab2xl procyear using /*redacted*/, col(1) row(1) replace

        sort sitepatientid siterecordid

        tab datecollectedprecision,m
        tab dateprocessedprecision,m
        count if datecollected==.
        count if dateprocessed==.
        tab colyear if dateprocessed==
        tab colyear if dateprocessed==.
    * /*redacted*/ specimens have no processed date, varying years no specific missing pattern

        tab specimentype,m
        tab anticoagulant,m
        tab additive,m
        tab specimenform,m
        tab numberofaliquots,m
        tab volumeperaliquot,m
        tab numberofsections,m
        tab numberofcells,m
        tab colyear if real(numberofcells)<0
    * negative number of cells?
        tab storagetemperature,m
        tab numberofaliquots specimentype,m

        gen duration=dateprocessed-datecollected if dateprocessed!=.
        tab duration,m
        list if duration<0
        codebook sitepatientid if duration>30 & dateprocessed!=.
    * processed date before collected date?
    restore
    clear



    ***  visitAppointment ***

    clear
    use /*redacted*/,clear
        encode encountertype, gen(type)
        gen year=yofd(encounterdate)
        tab type,m

        graph bar (count) type if encountertype=="Initial", over(year) blabel(bar)
        sort sitepatientid encounterdate encountertype
        tab apptstatus,m
        tab encountertype,m
        tab department,m
        replace department=trim(department)
        tab encounterlocation,m
        tab encountertype encounterlocation,m
        format encounterdate %tdCY-N-D


    preserve
        gen enrollyear=yofd(encounterdate) if encountertype=="Initial"
        tab enrollyear encountertype  if encountertype=="Initial"
        keep if encountertype=="Initial"
        sort enrollyear
        by enrollyear: gen count=_N
        duplicates drop enrollyear,force

        graph bar count, over(enrollyear) ytitle("Number of initial visits") yscale(nofextend)
    restore

    preserve
        drop  encounterinstype1 encounterinstype2 encounterinstype3 encounterinstype4 encounterinstype5 encounterid scheduledate
        by sitepatientid: gen rank=_n
        by sitepatientid: egen maxrank=max(rank)
        by sitepatientid: egen maxrank=max(rank)
        keep if maxrank==rank
        keep sitepatientid encounterdate encountertype
        sort sitepatientid
        save /*redacted*/, replace
    restore

    preserve
        drop  encounterinstype1 encounterinstype2 encounterinstype3 encounterinstype4 encounterinstype5 encounterid scheduledate
        keep if encountertype=="Initial"
        save /*redacted*/, replace
    restore

    preserve
        drop apptstatus encounterinstype1 encounterinstype2 encounterinstype3 encounterinstype4 encounterinstype5 encounterid scheduledate
        by sitepatientid, sort: egen ini=min(cond(encountertype=="Initial",encounterdate,.))
        by sitepatientid, sort: egen t1=min(cond(encountertype=="HIV primary care", encounterdate,.))
        format ini t1 %tdCY-N-D
        gen checkpatient=1 if (t1-ini)>365.25
        tab checkpatient,m
        capture: codebook sitepatientid if checkpatient==1
        * to check if any patients are ineligible for not having at least two primary care visits in a 12 month (365.25 days) period*/
        gen checkt1=1 if t1<ini
        tab checkt1,m /*to check if Initial is badly coded as evidenced by even one instance of HIV primary care being before Initial*/
        codebook sitepatientid if checkt1!=.
        gen daysto2ndPC=t1-ini if encountertype=="Initial"
        tab daysto2ndPC,m
        tab daysto2ndPC
    * how many days passed between a patient's initial PC visit and the second visit*/
    * note the pattern, weekly surge in multiples of 7 up to a maximun of 25 weeks (175 days) and then it disappears*/
    * possible reasons?*/
    * 95% of patients have a second visit within /*redacted*/ days*/
    * 63% within /*redacted*/ weeks*/
    restore

    preserve
        sort encounterdate
        gen rank=_n
        list if rank==1
        drop rank
        gsort -encounterdate
        gen rank=_n
        list if rank==1
        drop rank
    restore

    clear

log close DataReviewNovember2017


        ** LE FIN **
        * END OF CODE *
```