

```

1 version 13.1
2
3
4 log using /*redacted*/ , append smcl name(DataReviewNovember2017)
5 * Saurabh Chavan
6
7 * This code will examine the prepared data files for /*redacted*/ for
inconsistencies and incompatibilities with the /*redacted*/ data requirements
8
9 * Tables are examined in alphabetical order
10
11 clear
12 capture
13
14 *** /*redacted*/ ***
15
16 * We are not submitting /*redacted*/ table this time
17 * It could be a consideration for the March 2018 or July 2018 upload after a review
of /*redacted*/ and /*redacted*/ data
18
19 /*use /*redacted*/
20
21 sort sitepatientid
22 tab datagroup,m
23
24 clear
25 capture
26 */
27
28 *** /*redacted*/ ***
29
30 * this table can be used to delete erroneous records that were uploaded earlier
31
32 use /*redacted*/ , clear
33 tab recordtype,m
34
35
36 clear
37 capture
38
39 *** /*redacted*/ ***
40
41 use /*redacted*/
42 sort sitepatientid
43
44 * gender
45 tab presentsex birthsex,m row col
46 tab transgender,m
47 table presentsex birthsex, by(transgendered)
48 gen checkgender=1 if presentsex==birthsex & transgendered=="Yes" & presentsex!="
49 /* checking for inconsistent gender */
50 tab checkgender,m
51 drop checkgender
52
53 * death
54 tab deathdate if deathdate<(date("04-01-2000","MDY")) & deathdate>(date(
"10-31-2017","MDY")),m
55 /* to check implausible deathdates if before cohort start date or after database
cut date */
56 count if deathdate!=.
57 * /*redacted*/ deceased patients
58 tab deathdatesource deathdateprecision ,m

```

```

59     gen deathyear=yofd(deathdate)
60     tab deathdatesource,m
61
62     tab2xl deathyear using /*redacted*/, col(1) row(1) replace
63     histogram deathyear
64     tab deathyear, plot
65
66     * birthyear/age
67     tab2xl birthyear using /*redacted*/, col(1) row(1) replace
68     tab birthyear,m
69     gen age=2017-birthyear if deathdate==.
70     replace age=deathyear-birthyear if deathdate!=.
71     tab2xl age using /*redacted*/, col(1) row(1) replace
72     summarize age, detail
73     histogram age, normal
74
75     * race/ethnicity
76     tab race hispanic,m row col
77     list sitepatientid if race=="" & hispanic=="
78
79     preserve
80     keep if race=="" & hispanic=="
81     /* missing race/ethnicity */
82     destring sitepatientid, replace force
83     merge 1:1 sitepatientid using /*redacted*/
84     keep if _merge==3
85     keep mrn sitepatientid pat_lname pat_fname lastname firstname dob ssn newpatient
86     export excel using /*redacted*/, sheet("missingracehisp") firstrow(var) replace
87     restore
88
89     tab birthcountry,m
90     drop age deathyear
91
92     clear
93
94     *** /*redacted*/ ***
95     clear
96
97     use /*redacted*/,clear
98
99     sort sitepatientid siterecordid
100
101     tab datasource,m
102     tab diagnosisdateprecision,m
103
104     codebook sitepatientid
105     * /*redacted*/ patients have any diagnosis
106
107     duplicates tag sitepatientid diagnosisname diagnosisdate, gen(dupdx)
108     /* duplicate diagnoses */
109     tab dupdx,m
110
111     preserve
112     keep if dupdx>0
113     gsort -dupdx sitepatientid diagnosisdate diagnosisname
114     capture: export excel using /*redacted*/, sheet("duplicatedx") firstrow(var)
sheetmodify
115     restore
116
117     drop dupdx
118     preserve
119
120     tab deathdateprecision, col(1) row(1) replace
121     histogram deathdateprecision
122     tab deathdateprecision, plot

```

```

120     gen dxproblemdate="checkfulldate" if diagnosisdateprecision=="unknown" &
diagnosisdate!=date("01/01/1900","MDY")
121     * this will mark all non "01/01/1900" dates that have unknown precision (which it
shouldn't be)
122
123     replace dxproblemdate="checkmonth" if diagnosisdateprecision=="year" & month(
diagnosisdate)!=1
124     * this will mark something like 04/01/2011 or 04/04/2011 since year precision is only
01/01/someyear and nothing else
125
126     replace dxproblemdate="checkday" if diagnosisdateprecision=="year" & day(
diagnosisdate)!=1
127     * this will mark something like 01/04/2011 or 04/04/2011 since year precision is only
01/01/someyear and nothing else
128
129     replace dxproblemdate="checkdaymonth" if diagnosisdateprecision=="year" & day(
diagnosisdate)!=1 & month(diagnosisdate)!=1
130     * this will remark 04/04/2011 from above since month precision is always
somemonth/01/someyear and nothing else
131
132     replace dxproblemdate="checkday" if diagnosisdateprecision=="month" & day(
diagnosisdate)!=1
133     keep if dxproblemdate!=""
134
135     capture: export excel using /*redacted*/, sheet("dxproblemdate") firstrow(var)
sheetmodify
136     save /*redacted*/, replace
137     restore,preserve
138
139     gen dxdateCR="01jananyyear" if diagnosisdateprecision=="year" & day(diagnosisdate
)==1 & month(diagnosisdate)==1
140     * this will mark something like jan/01/2000 to check if it is not underprecise (could
be month or day instead)
141
142     replace dxdateCR="01anymonanyyear" if diagnosisdateprecision=="month" & inlist(
month(diagnosisdate),1,2,3,4,5,6,7,8,9,10,11,12) & day(diagnosisdate)==1
143     * this will mark something like sep/01/1990 to check if it is not underprecise (could
be day instead)
144
145     replace dxdateCR="01anymonanyyear" if diagnosisdateprecision=="day" & inlist(month
(diagnosisdate),1,2,3,4,5,6,7,8,9,10,11,12) & day(diagnosisdate)==1
146     * this will mark something like sep/01/1990 to check if it is not overprecise (could
be month instead)
147
148     keep if dxdateCR!=""
149     sort sitepatientid diagnosisdate diagnosisname
150     capture: export excel using /*redacted*/, sheet("dxdatechartreview") firstrow(var)
sheetmodify
151
152     restore
153     preserve
154
155     merge m:1 sitepatientid using /*redacted*/
156     drop if _merge==2
157     drop _merge
158
159     gen dxafterlastvisit=1 if encounterdate<diagnosisdate & diagnosisdateprecision==
"day" & yofd(encounterdate)<2016
160
161     * there should not be any diagnoses after the last visit date if the visit is from
2016 or before
162     * is it possible to have a diagnosis date months after the last recorded visit?
163     * it is possible in 2017 if the diagnosis and visits table were not cut for the same

```

```

163  * It is possible in 2017 if the diagnosis and visits table were not cut for the same
last through date
164
165      tab dxafterlastvisit,m
166      sort dxafterlastvisit sitepatientid diagnosisdate diagnosisname
167      keep if dxafterlastvisit==1
168      keep siterecordid sitepatientid diagnosisname diagnosisdate diagnosisdateprecision
datasource historical encounterdate encountertype department encounterlocation
dxafterlastvisit
169      sort sitepatientid diagnosisdate diagnosisname encounterdate
170
171      capture: export excel using /*redacted*/, sheet("dxafterlastvisit") firstrow(var)
sheetmodify
172      restore
173      preserve
174
175      * there should be no diagnoses after deathdate and there aren't any
176
177      merge m:1 sitepatientid using /*redacted*/
178      keep if _merge==3
179      drop _merge
180      gen dataafterdeathdate=1 if (deathdate<diagnosisdate) & deathdate!=. &
diagnosisdate!=. & diagnosisdateprecision=="day" & diagnosisdate!=date("01/01/1900","MDY")
181      sort dataafterdeathdate
182      tab dataafterdeathdate,m
183      capture: export excel using /*redacted*/, firstrow(var) sheet("dxafterdeath")
sheetmodify
184      restore
185
186      gen dxyear=yofd(diagnosisdate) if yofd(diagnosisdate)!=1900
187      histogram dxyear
188      tab dxyear, plot
189
190      tab2x1 dxyear using /*redacted*/, row(1) col(1)
191      tab historical,m
192
193
194      * 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.
195      * /*redacted*/ wants all diagnosis dates but are they ok with more than two-three
rows per day?
196      * Confirm with /*redacted*/?
197
198
199      *** /*redacted*/ ***
200
201      clear
202      use /*redacted*/
203      codebook sitepatientid
204
205      preserve
206      duplicates drop sitepatientid testdate, force
207      sort sitepatientid testdate
208      by sitepatientid: gen count=_N
209      tab count,m
210      restore
211
212      tab mutation, sort
213      * capture top ten mutations
214
215
216      tab mutation if mutation=="NULL"
217      preserve

```

```

218     sort testdate
219     gen rank=_n
220     list if rank==1
221 restore,preserve
222     gsort -testdate
223     gen rank=_n
224     list if rank==1
225 restore
226
227     bysort sitepatientid testdate: gen testcount=_N
228     bysort sitepatientid testdate: gen testrank=_n
229     by sitepatientid: gen testyear=yofd(testdate) if testcount==testrank
230
231     tab2xl testyear using /*redacted*/, col(1) row(1) replace
232     tab testyear, plot
233
234     * first test in 2001-06-27
235     * last test in 2015-04-27
236
237     * none in 2016 - 2017?
238     * what about those that joined the cohort after April 2015?
239     * what about tests before June 2001? /*redacted*/ has said we let go of these
240     * prioritise for /*redacted*/; waiting to hear from data team
241
242
243     *** /*redacted*/ ***
244
245     clear
246     use /*redacted*/
247
248     codebook sitepatientid
249     codebook sitepatientid if zipcode=="ZZZZZ"
250
251     codebook sitepatientid if (real(zipcode)>94102 & real(zipcode)<94188) & real(
zipcode)!=.
252
253     preserve
254     keep if (real(zipcode)>94102 & real(zipcode)<94188) & real(zipcode)!=.
255     /* local zipcodes */
256     gen zip=real(zipcode)
257     duplicates drop sitepatientid zip, force
258
259     tab2xl zip using /*redacted*/, row(1) col(1) replace
260 restore
261
262     *** /*redacted*/ ***
263
264     clear
265     use /*redacted*/
266     * gen double adm=clock(admitdate,"MDYhms")
267     * gen double dsc=clock(dischargedate,"MDYhms")
268     * format adm %tc
269     * format dsc %tc
270     codebook sitepatientid
271
272     preserve
273     tab admitdateprecision,m
274     tab dischargedateprecision,m
275
276     sort sitepatientid admitdate
277     gen checkschdate=1 if dischargedate<admitdate
278     tab checkschdate,m

```

```

279     drop checkdschdate
280     gen longstay=">6mon" if ((dischargedate-admitdate)/30)>6
281     tab longstay,m
282     codebook sitepatientid if longstay!=""
283     * /*redacted*/ patients had /*redacted*/ admissions with stays longer than 6 months
284     gen hospstay=(dischargedate-admitdate)
285     gsort -hospstay
286     gsort sitepatientid -hospstay    longstay admitdate
287
288     sort admitdate
289     gen rank=_n
290     egen firstadm=min(rank) if rank==1
291     list sitepatientid siterecordid admitdate admitdateprecision dischargedate
dischargedateprecision if firstadm==1
292     gsort -admitdate
293     replace rank=_n
294     egen lastadm=min(rank) if rank==1
295     list sitepatientid siterecordid admitdate admitdateprecision dischargedate
dischargedateprecision if lastadm==1
296
297     sort dischargedate
298     replace rank=_n
299     egen firstdsc=min(rank) if rank==1
300     list sitepatientid siterecordid admitdate admitdateprecision dischargedate
dischargedateprecision if firstdsc==1
301     gsort -dischargedate
302     replace rank=_n
303     egen lastdsc=min(rank) if rank==1
304     list sitepatientid siterecordid admitdate admitdateprecision dischargedate
dischargedateprecision if lastdsc==1
305
306     drop rank
307     sort sitepatientid admitdate
308     by sitepatientid: gen rank=_n
309     summarize rank, detail
310     summarize hospstay, detail
311
312     gen admityear=yofd(admitdate)
313     tab admityear,m plot
314     tab2xl admityear using /*redacted*/, row(1) col(1) replace
315     bysort admityear: gen count=_N
316     duplicates drop admityear,force
317     keep admityear count
318
319     * histogram of admissions by year
320
321     graph bar (sum) count, over(admityear)
322
323     restore
324
325     * no observations from 2000,2001?? /*redacted*/ said this is all right. We let it go
326
327
328     *** /*redacted*/ ***
329
330     * why are all stop dates unknown???
331     * we are submitting only one insurance type for each patient, that too the most recent
332     * /*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
333
334     clear

```

```

335 use /*redacted*/
336 * gen int insstodate=dofc(clock(insurancstartdate,"MDYhms"))
337 * format insstodate %td
338 * gen int insstodate=dofc(clock(insurancstopdate,"MDYhms"))
339 * format insstodate %td
340 sort sitepatientid insurancstartdate insurancetype
341
342 tab insurance,m
343 tab insurancstartdateprecision insurancstopdateprecision,m
344 table insurancstartdateprecision insurancstopdateprecision,m by(insurancetype)
345
346 clear
347
348 *** /*redacted*/ ***
349
350 use /*redacted*/
351
352
353 * gen double resdatetime=clock(resultdate,"MDYhms")
354 * format resdatetime %tc
355 * gen resdate=dofc(clock(resultdate,"MDYhms"))
356 * format resdate %td
357
358
359 * to check the first and last lab dates*/
360 sort resultdate
361 gen rank=_n
362 egen firstlab=min(rank) if resultdate!=date("01/01/1900","MDY")
363 egen lastlab=max(rank)
364 list if firstlab==rank | lastlab==rank
365 drop firstlab lastlab rank
366
367 preserve
368 gen labproblemdate="checkfulldate" if resultdateprecision=="unknown" & resultdate
!=date("01/01/1900","MDY")
369 * this will mark all non "01/01/1900" dates that have unknown precision (which it
shouldn't be)
370
371 replace labproblemdate="checkmonth" if resultdateprecision=="year" & month(
resultdate)!=1
372 * this will mark something like 04/01/2011 or 04/04/2011 since year precision is only
01/01/someyear and nothing else
373
374 replace labproblemdate="checkday" if resultdateprecision=="year" & day(resultdate
)!=1
375 * this will mark something like 01/04/2011 or 04/04/2011 since year precision is only
01/01/someyear and nothing else
376
377 replace labproblemdate="checkdaymonth" if resultdateprecision=="year" & day(
resultdate)!=1 & month(resultdate)!=1
378 * this will remark 04/04/2011 from above since month precision is always
somemonth/01/someyear and nothing else
379
380 replace labproblemdate="checkday" if resultdateprecision=="month" & day(resultdate
)!=1
381 keep if labproblemdate!=""
382 tab labproblemdate,m
383 capture: export excel using /*redacted*/, sheet("labdatechartreview1") firstrow(
var) sheetmodify
384 restore,preserve
385 gen resdateCR="01jananyyear" if resultdateprecision=="year" & day(resultdate)==1 &
month(resultdate)==1

```

```

386     * this will mark something like jan/01/2000 to check if it is not underprecise (could
    be month or day instead)
387
388     replace resdateCR="01anymonanyyear" if resultdateprecision=="month" & inlist(month
(resultdate),1,2,3,4,5,6,7,8,9,10,11,12) & day(resultdate)==1
389     * this will mark something like sep/01/1990 to check if it is not underprecise (could
    be day instead)
390
391     replace resdateCR="01anymonanyyear" if resultdateprecision=="day" & inlist(month(
resultdate),1,2,3,4,5,6,7,8,9,10,11,12) & day(resultdate)==1
392     * this will mark something like sep/01/1990 to check if it is not overprecise (could
    be month instead)
393
394     keep if resdateCR!=""
395     tab resdateCR,m
396     capture: export excel using /*redacted*/, sheet("labdatechartreview2") firstrow(
var) sheetmodify
397     restore,preserve
398
399     * these issues above were not observed
400
401     encode testname, gen(name)
402     labelbook name
403     tab name,m
404     drop name
405     codebook testname
406     * trailing blanks in many responses*/
407     tab result if testname=="Height",m
408     tab result if testname=="Weight",m
409     tab result if testname=="CD4 cell absolute",m
410
411     tab units,m
412     * are non uniform units ok?
413     egen minmax=concat(normalmin normalmax unit testname),punct(" ")
414     table minmax,m
415     drop minmax
416     tab historical,m
417     restore,preserve
418     merge m:1 sitepatientid using /*redacted*/
419     gen dataafterdeathdate=1 if (deathdate<dofc(resultdate)) & deathdate!=. &
resultdate!=. & resultdateprecision=="day" & dofc(resultdate)!=date("01/01/1900","MDY")
420     tab dataafterdeathdate _merge,m
421     restore,preserve
422     * to identify potential misclassification of the data source
423     * if precision is time then how come data source is "Source unknown"?
424
425     tab datasource resultdateprecision,m
426     codebook resultdate if datasource=="Source unknown" & resultdateprecision=="time"
427     gen checksource=1 if datasource=="Source unknown" & resultdateprecision=="time"
428     tab checksource,m
429     keep if checksource!=.
430     keep sitepatientid result resultdate testname
431     capture: export excel using /*redacted*/, sheet("labsourceCR") firstrow(var)
sheetmodify
432     restore,preserve
433     codebook sitepatientid if strmatch(testname,"*HIV*")==1
434     codebook sitepatientid
435     * /*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*/)
436     * what to do if patients qualify for /*redacted*/ but have not had HIV test at
/*redacted*/ for months?

```



```

437  /*redacted*/ 101 months?
438      codebook sitepatientid if strmatch(testname,"*CD4 cell absolute*")==1
439      codebook sitepatientid if strmatch(testname,"*HIV-1 RNA*")==1 | strmatch(testname,
440  /*redacted*/ have at least one CD4 count
441      sort sitepatientid resultdate
442      restore
443
444      preserve
445      keep if strmatch(testname,"*CD4 cell absolute*")==1
446      destring result, replace force
447      sort sitepatientid resultdate
448      by sitepatientid: gen rank=_n
449      keep if rank==1
450      summarize result, detail
451      restore,preserve
452      keep if strmatch(testname,"*HIV-1 RNA*")==1 | strmatch(testname,"*HIV-1 Viral*")==1
453      replace result = regexs(0) if regexm(result, "[0-9]*$") /*to remove < > = from
the results*/
454      destring result, replace force
455      sort sitepatientid resultdate
456      by sitepatientid: gen rank=_n
457      keep if rank==1
458      summarize result, detail
459      restore
460
461      clear
462
463      *** /*redacted*/ ***
464
465      clear
466
467      use /*redacted*/, clear
468      preserve
469      duplicates tag sitepatientid startdate medicationname, gen (duplicatemeds)
470      tab duplicatemeds,m
471      sort sitepatientid startdate medicationname
472      sort medicationname
473      merge m:1 medicationname using /*redacted*/.dta
474      drop if _merge==2
475      destring sitepatientid, replace force
476      sort sitepatientid startdate medicationname
477      by sitepatientid startdate: gen artcount=1 if code=="ART"
478      by sitepatientid startdate: egen maxart=sum(artcount) if startdate!=date(
479  "01/01/1900","MDY") | startdate!=enddate
480      tab maxart,m /* maximum number of ARTs prescribed */
481      keep maxart sitepatientid medicationname startdate startdateprecision enddate
enddateprecision
482      gsort -maxart sitepatientid startdate medicationname
483      keep if maxart!=.
484      capture: export excel using /*redacted*/, sheet("maxART") firstrow(var) sheetmodify
485      restore
486      preserve
487      gen checkenddate=0
488      replace checkenddate=1 if enddate < startdate & enddate!=date("01/01/1900","MDY")
& enddateprecision=="day"
489      tab checkenddate,m /*this will tag erroneous end dates that are earlier than
startdates*/
490      foreach i in bcheckenddate {
491      drop if `i'==0
492      sort sitepatientid startdate

```

```

502     sort sitepatientid startdate
493     save /*redacted*/, replace
494     capture: export excel using /*redacted*/, sheet("checkenddatemeds") firstrow(
var) sheetmodify
495     restore,preserve
496 }
497     restore
498     preserve
499     * to spot bad start dates*/
500     gen sproblemdate="checkfulldate" if startdateprecision=="unknown" & startdate!=
date("01/01/1900","MDY")
501     * this will mark all non "01/01/1900" dates that have unknown precision (which it
shouldn't be)
502
503     replace sproblemdate="checkmonth" if startdateprecision=="year" & month(startdate
)!=1
504     * this will mark something like 04/01/2011 or 04/04/2011 since year precision is only
01/01/someyear and nothing else
505
506     replace sproblemdate="checkday" if startdateprecision=="year" & day(startdate)!=1
507     * this will mark something like 01/04/2011 or 04/04/2011 since year precision is only
01/01/someyear and nothing else
508
509     replace sproblemdate="checkdaymonth" if startdateprecision=="year" & day(startdate
)!=1 & month(startdate)!=1
510     * this will remark 04/04/2011 from above since month precision is always
somemonth/01/someyear and nothing else
511
512     replace sproblemdate="checkday" if startdateprecision=="month" & day(startdate)!=1
513     tab sproblemdate,m
514     sort startdate sitepatientid
515     foreach name in sproblemdate {
516         drop if sproblemdate==" "
517         sort sproblemdate startdate sitepatientid
518         save /*redacted*/, replace
519         capture: export excel using /*redacted*/, sheet("problemprecisionstartmeds")
firstrow(var) sheetmodify
520         restore, preserve
521     }
522     restore
523     preserve
524     gen sdateCR="01jananyyear" if startdateprecision=="year" & day(startdate)==1 &
month(startdate)==1
525     * this will mark something like jan/01/2000 to check if it is not underprecise
(could be month or day instead)
526
527     replace sdateCR="01anymonanyyear" if startdateprecision=="month" & inlist(month(
startdate),1,2,3,4,5,6,7,8,9,10,11,12) & day(startdate)==1
528     * this will mark something like sep/01/1990 to check if it is not underprecise
(could be day instead)
529
530     replace sdateCR="01anymonanyyear" if startdateprecision=="day" & inlist(month(
startdate),1,2,3,4,5,6,7,8,9,10,11,12) & day(startdate)==1
531     * this will mark something like sep/01/1990 to check if it is not overprecise
(could be month instead)
532
533     tab sdateCR,m
534     sort startdate sitepatientid
535     foreach name in sdateCR {
536         drop if sdateCR==" "
537         sort sdateCR startdate sitepatientid
538         save /*redacted*/, replace
539         capture: export excel using /*redacted*/, sheet("medstartdateCR") firstrow(var)

```

```

sheetmodify
540     restore, preserve
541 }
542     restore
543     preserve
544     * to spot bad end dates*/
545     gen eproblemdate="checkfulldate" if enddateprecision=="unknown" & enddate!=date(
"01/01/1900","MDY")
546     * this will mark all non "01/01/1900" dates that have unknown precision (which it
shouldn't be)
547
548     replace eproblemdate="checkmonth" if enddateprecision=="year" & month(enddate)!=1
549     * this will mark something like 04/01/2011 or 04/04/2011 since year precision is only
01/01/someyear and nothing else
550
551     replace eproblemdate="checkday" if enddateprecision=="year" & day(enddate)!=1
552     * this will mark something like 01/04/2011 or 04/04/2011 since year precision is only
01/01/someyear and nothing else
553
554     replace eproblemdate="checkdaymonth" if enddateprecision=="year" & day(enddate)!=1
& month(enddate)!=1
555     * this will remark 04/04/2011 from above since month precision is always
somemonth/01/someyear and nothing else
556     replace eproblemdate="checkday" if enddateprecision=="month" & day(enddate)!=1
557     tab eproblemdate,m
558     sort startdate sitepatientid
559     foreach name in eproblemdate {
560         drop if eproblemdate==" "
561         sort eproblemdate startdate sitepatientid
562         save /*redacted*/, replace
563         capture: export excel using /*redacted*/, sheet("problemprecisionendmeds") firstrow(
var) sheetmodify
564         restore, preserve
565     }
566     restore
567     preserve
568     gen edateCR="01jananyyear" if enddateprecision=="year" & day(enddate)==1 & month(
enddate)==1
569     * this will mark something like jan/01/2000 to check if it is not underprecise (could
be month or day instead)
570
571     replace edateCR="01anymonanyyear" if enddateprecision=="month" & inlist(month(
enddate),1,2,3,4,5,6,7,8,9,10,11,12) & day(enddate)==1
572     * this will mark something like sep/01/1990 to check if it is not underprecise (could
be day instead)
573
574     replace edateCR="01anymonanyyear" if enddateprecision=="day" & inlist(month(
enddate),1,2,3,4,5,6,7,8,9,10,11,12) & day(enddate)==1
575     * this will mark something like sep/01/1990 to check if it is not overprecise (could
be month instead)
576
577     tab edateCR,m
578     sort startdate sitepatientid
579     foreach name in edateCR {
580         drop if edateCR==" "
581         sort edateCR startdate sitepatientid
582         save /*redacted*/, replace
583         capture: export excel using /*redacted*/, sheet("medenddateCR") firstrow(var)
sheetmodify
584         restore, preserve
585     }
586     restore

```

```

587
588     preserve
589     * to identify startdates==enddates
590     gen samedate="N0"
591     replace samedate="YES" if startdate==enddate & startdateprecision=="day" &
enddateprecision=="day"
592     replace samedate="OK ongoing" if samedate=="YES" & enddatetype=="Ongoing"
593     replace samedate="OK statmed" if samedate=="YES" & sig=="Stat"
594
595     table enddatetype samedate, by(datasource)
596     sort samedate startdate startdateprecision
597     * this will give an idea of how these three interrelate and should raise appropriate
suspicions for certain combinations
598     restore
599
600     preserve
601     duplicates drop medicationname,force
602     sort medicationname
603     gen rank=_n
604     keep medicationname rank
605     save /*redacted*/,replace
606     clear
607     import excel using /*redacted*/, sheet("standardCodes_Medication") firstrow
608     sort code
609     rename code medicationname
610     merge medicationname using medicationname
611     tab _merge
612     keep category medicationname _merge
613     gen medstatus="in CNICS" if _merge==1
614     replace medstatus="in Upload" if _merge==2
615     replace medstatus="in both" if _merge==3
616     drop _merge
617     sort medstatus category medicationname
618     save /*redacted*/,replace
619
620     restore
621     * to indirectly confirm whether or not all CNICS required medications are covered
622
623     tab form route,m
624     * to identify implausible combinations like injectable pills*/
625     preserve
626     gen checkformroute=0
627     replace checkformroute=1 if route=="IM" & inlist(form,"injection","injectable
solution","solution")==0
628     replace checkformroute=1 if route=="IV" & inlist(form,"injection","injectable
solution","solution")==0
629     replace checkformroute=1 if route=="PO" & inlist(form,"pill","troches","syrup",
"tablet","capsule")==0
630     replace checkformroute=1 if route=="PR" & inlist(form,"suppository")==0
631     replace checkformroute=1 if route=="PV" & inlist(form,"cream","suppository")==0
632     replace checkformroute=1 if route=="SL" & inlist(form,"pill")==0
633     replace checkformroute=1 if route=="inhalation" & inlist(form,"puff","inhaler",
"liquid")==0
634     replace checkformroute=1 if route=="intranasal" & inlist(form,"puff")==0
635     replace checkformroute=1 if inlist(route,"intraocular (right)","intraocular",
"intraocular (left)")==1 & inlist(form,"solution")==0
636     replace checkformroute=1 if route=="sub-Q" & inlist(form,"injection","injectable
solution")==0
637     replace checkformroute=1 if route=="topical" & inlist(form,"cream","gel","liquid",
"lotion","ointment","patch","solution")==0
638
639     * since route and form both may be missclassified, one needs to check what the

```

```

640 medication is in order to make the correct determination for both
641     tab checkformroute,m
642     tab strength,m
643     tab units,m
644     tab sig,m
645     tab enddatetype,m
646     tab enddateprecision,m
647     tab datasource,m
648     tab stopreason,m
649     tab historical,m
650
651 * to see the first and the last medication prescribed
652     sort startdate startdateprecision
653     gen rank=_n
654     egen firstmed=min(rank) if startdate!=date("01/01/1900","MDY")
655     list sitepatientid medicationname startdate startdateprecision enddate
656     enddateprecision enddatetype datasource if rank==firstmed
657
658     gsort -startdate
659     replace rank=_n
660     egen lastmed=min(rank) if rank==1
661     list sitepatientid medicationname startdate startdateprecision enddate
662     enddateprecision enddatetype datasource if lastmed==1
663
664     restore
665     clear
666
667 *** /*redacted*/ ***
668 * we have not submitted this table traditionally. we submit the medication table
669
670 *** /*redacted*/ ***
671 use /*redacted*/
672     tab CODcodeLabel,m
673     tab type,m
674     tab source,m
675
676 *** pro ***
677
678 * as this table is generated through a separate and independent automated process,
679 there isn't much scope for inconsistencies to creep in as such
680
681 clear
682 use /*redacted*/
683     sort sitepatientid siterecordid
684     tab projectid,m
685     tab sessionid,m
686     tab questionid,m
687     tab sequence,m
688     tab state,m
689     tab value,m
690
691 clear
692
693 *** procedure ***
694
695 use /*redacted*/ ,clear
696     tab siteprocedure,m
697     codebook sitepatientid
698
699 clear
700
701 *** /*redacted*/ ***

```

```

698 * two possible inconsistencies in this table are duplicate risks and incorrectly
699 coded risks
700 clear
701 use /*redacted*/,clear
702 preserve
703 sort sitepatientid siterecordid
704 encode risk, gen(risktype)
705 labelbook risktype
706 tab risk,m
707
708 * check if properly coded as /*redacted*/
709 duplicates tag sitepatientid risk, gen(duprisk)
710 tab duprisk,m
711
712 bysort sitepatientid: gen riskcount=_n
713 /* maximum number of discrete risks */
714 bysort sitepatientid: egen maxrisk=max(riskcount)
715
716 count if riskcount==maxrisk & riskcount==1
717 count if riskcount==maxrisk & riskcount==2
718 count if riskcount==maxrisk & riskcount==3
719 count if riskcount==maxrisk & riskcount==4
720
721 codebook sitepatientid
722 * /*redacted*/ of /*redacted*/ have risks recorded - please see /*redacted*/
723 restore
724
725 clear
726
727 *** specimenTracking ***
728 clear
729
730 use /*redacted*/,clear
731 gen datecol=date(datecollected,"MDY")
732 drop datecollected
733 rename datecol datecollected
734 gen dateproc=date(dateprocessed,"MDY")
735 drop dateprocessed
736 rename dateproc dateprocessed
737 format datecollected dateprocessed %tdCY-N-D
738 preserve
739 sort datecollected
740 gen rank=_n
741 list if rank==1
742 gsort -datecollected
743 replace rank=_n
744 list if rank==1
745 drop rank
746 gen colyear=year(datecollected)
747 gen procyar=year(dateprocessed)
748 tab colyear,m
749 tab procyar,m
750 tab2xl colyear using /*redacted*/, col(1) row(1) replace
751 tab2xl procyar using /*redacted*/, col(1) row(1) replace
752
753 sort sitepatientid siterecordid
754
755 tab datecollectedprecision,m
756 tab dateprocessedprecision,m
757 count if datecollected==.
758 count if dateprocessed==.
759 tab colyear if dateprocessed==.

```

```

759     tab colyear if dateprocessed==.
760     * /*redacted*/ specimens have no processed date, varying years no specific missing
pattern
761
762     tab specimentype,m
763     tab anticoagulant,m
764     tab additive,m
765     tab specimenform,m
766     tab numberofaliquots,m
767     tab volumeperaliquot,m
768     tab numberofsections,m
769     tab numberofcells,m
770     tab colyear if real(numberofcells)<0
771     * negative number of cells?
772     tab storagetemperature,m
773     tab numberofaliquots specimentype,m
774
775     gen duration=dateprocessed-datecollected if dateprocessed!=.
776     tab duration,m
777     list if duration<0
778     codebook sitepatientid if duration>30 & dateprocessed!=.
779     * processed date before collected date?
780     restore
781     clear
782
783
784
785     *** visitAppointment ***
786
787     clear
788     use /*redacted*/,clear
789     encode encountertype, gen(type)
790     gen year=yofd(encounterdate)
791     tab type,m
792
793     graph bar (count) type if encountertype=="Initial", over(year) blabel(bar)
794     sort sitepatientid encounterdate encountertype
795     tab apptstatus,m
796     tab encountertype,m
797     tab department,m
798     replace department=trim(department)
799     tab encounterlocation,m
800     tab encountertype encounterlocation,m
801     format encounterdate %tdCY-N-D
802
803
804     preserve
805     gen enrollyear=yofd(encounterdate) if encountertype=="Initial"
806     tab enrollyear encountertype if encountertype=="Initial"
807     keep if encountertype=="Initial"
808     sort enrollyear
809     by enrollyear: gen count=_N
810     duplicates drop enrollyear,force
811
812     graph bar count, over(enrollyear) ytitle("Number of initial visits") yscale(
nofextend)
813     restore
814
815     preserve
816     drop encounterinstype1 encounterinstype2 encounterinstype3 encounterinstype4
encounterinstype5 encounterid scheduledate
817     by sitepatientid: gen rank=_n
818     by sitepatientid: egen maxrank=max(rank)

```

```

819     keep if maxrank==rank
820     keep sitepatientid encounterdate encountertype
821     sort sitepatientid
822     save /*redacted*/, replace
823     restore
824
825     preserve
826     drop encounterinstype1 encounterinstype2 encounterinstype3 encounterinstype4
827     drop encounterinstype5 encounterid scheduledate
828     keep if encountertype=="Initial"
829     save /*redacted*/, replace
830     restore
831
832     preserve
833     drop apptstatus encounterinstype1 encounterinstype2 encounterinstype3
834     drop encounterinstype4 encounterinstype5 encounterid scheduledate
835     by sitepatientid, sort: egen ini=min(cond(encountertype=="Initial",encounterdate
836     ,.))
837     by sitepatientid, sort: egen t1=min(cond(encountertype=="HIV primary care",
838     encounterdate,.))
839     format ini t1 %tdCY-N-D
840     gen checkpatient=1 if (t1-ini)>365.25
841     tab checkpatient,m
842     capture: codebook sitepatientid if checkpatient==1
843     * to check if any patients are ineligible for not having at least two primary
844     care visits in a 12 month (365.25 days) period*/
845     gen checkt1=1 if t1<ini
846     tab checkt1,m /*to check if Initial is badly coded as evidenced by even one
847     instance of HIV primary care being before Initial*/
848     codebook sitepatientid if checkt1!=.
849     gen daysto2ndPC=t1-ini if encountertype=="Initial"
850     tab daysto2ndPC,m
851     tab daysto2ndPC
852     * how many days passed between a patient's initial PC visit and the second visit*/
853     * note the pattern, weekly surge in multiples of 7 up to a maximum of 25 weeks (175
854     days) and then it disappears*/
855     * possible reasons?*/
856     * 95% of patients have a second visit within /*redacted*/ days*/
857     * 63% within /*redacted*/ weeks*/
858     restore
859
860     preserve
861     sort encounterdate
862     gen rank=_n
863     list if rank==1
864     drop rank
865     gsort -encounterdate
866     gen rank=_n
867     list if rank==1
868     drop rank
869     restore
870
871     clear
872
873     log close DataReviewNovember2017
874
875     ** LE FIN **
876     * END OF CODE *

```