QC Findings

Things to Do

Please add any discrepencies you found between the original ADaM datasets from the CDISC Pilot and the ones we've programmed in R below.

ADSL

The R-generated ADSL matches the original ADSL from CDISC pilot data, besides the following mismatches: * Subject 01-702-1082 has a missing value for BMIBLGR1 in the R-generated ADSL, whilst BMIBLGR1 = "<25" in the original ADSL. This is an issue with the original ADSL, as this subject's BMI at baseline (BMIBL) is missing and therefore the subject shouldn't be assigned a BMI at baseline group.

ADAE

The R-generated ADAE matches the original ADAE from CDISC pilot data, besides the following mismatches: There is an issue with the original CDISC pilot dataset. ADURN is blank, where AESEQ is (5, 6, 7, 8) for the original CDISC dataset for Subject below:

```
> adae_orig %>% filter(USUBJID=='01-716-1418') %>% select(USUBJID,TRTSDT,ASTDT,AENDT,ADURN,ADURU,AESEQ)
# A tibble: 10 × 7
   USUBJID     TRTSDT     ASTDT      AENDT      ADURN ADURU AESEQ
   <chr>       <date>     <date>     <date>     <dbl> <chr> <dbl>
 1 01-716-1418 2013-05-05 2013-05-05 2013-05-07     3 DAY       1
 2 01-716-1418 2013-05-05 2013-05-05 NA            NA NA        2
 3 01-716-1418 2013-05-05 2013-05-05 2013-05-07     3 DAY       3
 4 01-716-1418 2013-05-05 2013-05-07 NA            NA NA        4
 5 01-716-1418 2013-05-05 2013-07-01 2013-09-26    NA NA        5
 6 01-716-1418 2013-05-05 2013-07-01 2013-10-04    NA NA        6
 7 01-716-1418 2013-05-05 2013-07-01 2013-09-26    NA NA        7
 8 01-716-1418 2013-05-05 2013-07-01 2013-10-04    NA NA        8
 9 01-716-1418 2013-05-05 2013-09-26 2013-11-11    47 DAY       9
10 01-716-1418 2013-05-05 2013-09-26 2013-11-11    47 DAY      10
```

Because it seems the original SDTM.AE.AESTDTC was missing Day, where it seems the original ADAE derivation for ADURN was probably using this date instead of the imputed date. Because day is missing in AESTDTC, ADURN can't derive days.

```
> ae %>% filter(USUBJID=='01-716-1418') %>% select(USUBJID,AESTDTC,AESEQ) %>% arrange(AESEQ)
# A tibble: 10 × 3
   USUBJID     AESTDTC    AESEQ
   <chr>       <chr>      <dbl>
 1 01-716-1418 2013-05-05     1
 2 01-716-1418 2013-05-05     2
 3 01-716-1418 2013-05-05     3
 4 01-716-1418 2013-05-07     4
 5 01-716-1418 2013-07        5
 6 01-716-1418 2013-07        6
 7 01-716-1418 2013-07        7
 8 01-716-1418 2013-07        8
 9 01-716-1418 2013-09-26     9
10 01-716-1418 2013-09-26    10
```

but the same records, derived in the Pilot 3 dataset do show a calculation since we are using the imputed

ASTDT, per the define (ADURN=AENDT-ASTDT+1).

```
#AE.AESTDTC, converted to a numeric SAS date. Some events with partial dates are imputed in a conservat
#value of '01' is used. If both the month and day are missing no imputation is performed as these dates
#of treatment. There are no events with completely missing start dates.

> adae0 %>% filter(USUBJID=='01-716-1418') %>% select(USUBJID,TRTSDT,ASTDT,AESTDTC,AENDT,AEENDY,ADURN,AD
# A tibble: 10 × 9
   USUBJID     TRTSDT     ASTDT      AESTDTC    AENDT      AEENDY ADURN ADURU AESEQ
   <chr>       <date>     <date>     <chr>      <date>      <dbl> <dbl> <chr> <dbl>
 1 01-716-1418 2013-05-05 2013-05-05 2013-05-05 2013-05-07      3     3 DAY       1
 2 01-716-1418 2013-05-05 2013-05-05 2013-05-05 NA             NA    NA NA        2
 3 01-716-1418 2013-05-05 2013-05-05 2013-05-05 2013-05-07      3     3 DAY       3
 4 01-716-1418 2013-05-05 2013-05-07 2013-05-07 NA             NA    NA NA        4
 5 01-716-1418 2013-05-05 2013-07-01 2013-07    2013-10-04    153    96 DAY       6
 6 01-716-1418 2013-05-05 2013-07-01 2013-07    2013-10-04    153    96 DAY       8
 7 01-716-1418 2013-05-05 2013-07-01 2013-07    2013-09-26    145    88 DAY       5
 8 01-716-1418 2013-05-05 2013-07-01 2013-07    2013-09-26    145    88 DAY       7
 9 01-716-1418 2013-05-05 2013-09-26 2013-09-26 2013-11-11    191    47 DAY       9
10 01-716-1418 2013-05-05 2013-09-26 2013-09-26 2013-11-11    191    47 DAY      10
```

This latter approach should be the correct approach.

Due to this, we have outlined the expected differences here :

```
> diffdf(adae, adae_orig, keys = c("STUDYID", "USUBJID", "AESEQ"))
Differences found between the objects!

A summary is given below.

There are columns in BASE and COMPARE with differing attributes !!
All rows are shown in table below
```

1. ADURN values will be populated in Pilot 3 (i.e. under `BASE`), following the latter derivation approach (i.e. ADURN=AENDT-ASTDT+1) for Subject 01-716-1418 where AESEQ is (5, 6, 7, 8) specified in define.

```
All rows are shown in table below

  ============================================================
   VARIABLE    STUDYID       USUBJID     AESEQ  BASE  COMPARE
  ------------------------------------------------------------
    ADURN     CDISCPILOT01  01-716-1418     5     88    <NA>
    ADURN     CDISCPILOT01  01-716-1418     6     96    <NA>
    ADURN     CDISCPILOT01  01-716-1418     7     88    <NA>
    ADURN     CDISCPILOT01  01-716-1418     8     96    <NA>
  ------------------------------------------------------------
```

2. ADURU should be set to 'DAYS' (i.e. under `BASE`) instead of 'DAY' when ADURN is not missing. Updated in Pilot 3 define.

```
First 10 of 718 rows are shown in table below

  ============================================================
   VARIABLE    STUDYID       USUBJID     AESEQ  BASE  COMPARE
  ------------------------------------------------------------
    ADURU     CDISCPILOT01  01-701-1015     3    DAYS    DAY
    ADURU     CDISCPILOT01  01-701-1023     1    DAYS    DAY
```

```
   ADURU    CDISCPILOT01  01-701-1023    4    DAYS    DAY
   ADURU    CDISCPILOT01  01-701-1047    1    DAYS    DAY
   ADURU    CDISCPILOT01  01-701-1047    2    DAYS    DAY
   ADURU    CDISCPILOT01  01-701-1097    2    DAYS    DAY
   ADURU    CDISCPILOT01  01-701-1097    3    DAYS    DAY
   ADURU    CDISCPILOT01  01-701-1097    5    DAYS    DAY
   ADURU    CDISCPILOT01  01-701-1097    6    DAYS    DAY
   ADURU    CDISCPILOT01  01-701-1097    7    DAYS    DAY
  ------------------------------------------------------------
```

ADLBC

The R-generated ADLBC matches the original ADLBC from CDISC pilot data, besides the following mismatches:

Three variables from R-generated ADLBC have class date while the same variables are numeric in the CDISC ADLBC. We opted to keep the date class in our R-generated ADLB.

```
> diffdf(adlbc, qc_adlbc, keys = c("STUDYID", "USUBJID", "AVISIT", "LBSEQ"))
Differences found between the objects!

A summary is given below.

There are columns in BASE and COMPARE with different classes !!
All rows are shown in table below

  ==================================
   VARIABLE  CLASS.BASE  CLASS.COMP
  ----------------------------------
     ADT        Date      numeric
    TRTEDT      Date      numeric
    TRTSDT      Date      numeric
  ----------------------------------
```

ADADAS

The R-generated ADADAS matches original adadas from CDISC pilot data, except for the records where `PARAMCD=ACTOT, DTYPE=LOCF`. This is an issue from the CDISC ADADAS.

- CDISC SDTM/qs: 818 records for `QSTESTCD=ACTOT`
- CDISC ADaM/adadas: 1040 records for `PARAMCD=ACTOT`, 799 (directly from qs, **should be 818**) + 241 imputed records (`DTYPE=LOCF`)
- adadas generated by R: 1040 records for `PARAMCD=ACTOT`, 818 (directly from qs) + 222 imputed records (`DTYPE=LOCF`)

Take a detailed example USUBJID="01-701-1294"

CDISC qs:

```
> qs %>% filter(QSTESTCD=="ACTOT") %>%
+   select(USUBJID, QSSEQ, VISIT, QSTESTCD, QSTEST,QSSTRESN) %>%
+   filter(USUBJID=="01-701-1294")
# A tibble: 4 × 6
  USUBJID     QSSEQ VISIT     QSTESTCD QSTEST                QSSTRESN
  <chr>       <dbl> <chr>     <chr>    <chr>                    <dbl>
```

```
1 01-701-1294  5015 BASELINE  ACTOT    ADAS-COG(11) Subscore        9
2 01-701-1294  5030 WEEK 8    ACTOT    ADAS-COG(11) Subscore       14
3 01-701-1294  5045 WEEK 12   ACTOT    ADAS-COG(11) Subscore        6
4 01-701-1294  5060 RETRIEVAL ACTOT    ADAS-COG(11) Subscore        9
```

CDISC adadas:

For the record with QSSEQ=5045 and AVISIT=Week 8, DTYPE is populated as LOCF , but this record is directly from qs dataset, not imputed.

```
> qc_adadas %>% filter(PARAMCD=="ACTOT") %>%
+   select(USUBJID, QSSEQ, PARAMCD, AVISITN, AVISIT, VISIT, AVAL, DTYPE, ANL01FL, ADT, ADY) %>%
+   arrange(USUBJID, AVISITN) %>% filter(USUBJID=="01-701-1294")
# A tibble: 5 × 11
  USUBJID     QSSEQ PARAMCD AVISITN AVISIT   VISIT      AVAL DTYPE  ANL01FL ADT          ADY
  <chr>       <dbl> <chr>     <dbl> <chr>    <chr>     <dbl> <chr>  <chr>   <date>     <dbl>
1 01-701-1294  5015 ACTOT         0 Baseline BASELINE      9 ""     "Y"     2013-03-24     1
2 01-701-1294  5030 ACTOT         8 Week 8   WEEK 8       14 ""     "Y"     2013-05-22    60
3 01-701-1294  5045 ACTOT         8 Week 8   WEEK 12      14 "LOCF" ""      2013-06-14    83
4 01-701-1294  5045 ACTOT        16 Week 16  WEEK 12      14 "LOCF" "Y"     2013-06-14    83
5 01-701-1294  5060 ACTOT        24 Week 24  RETRIEVAL     9 ""     "Y"     2013-10-08   199
```

adadas generated by R:

DTYPE is not LOCF for the record with QSSEQ=5045 and AVISIT=Week 8, as this record is directly from qs.

```
> adadas %>% filter(PARAMCD=="ACTOT") %>%
+   select(USUBJID, QSSEQ, PARAMCD, AVISITN, AVISIT, VISIT, AVAL, DTYPE, ANL01FL, ADT, ADY) %>%
+   arrange(USUBJID, AVISITN) %>% filter(USUBJID=="01-701-1294")
# A tibble: 5 × 11
  USUBJID     QSSEQ PARAMCD AVISITN AVISIT   VISIT      AVAL DTYPE  ANL01FL ADT          ADY
  <chr>       <dbl> <chr>     <dbl> <chr>    <chr>     <dbl> <chr>  <chr>   <date>     <dbl>
1 01-701-1294  5015 ACTOT         0 Baseline BASELINE      9 ""     "Y"     2013-03-24     1
2 01-701-1294  5030 ACTOT         8 Week 8   WEEK 8       14 ""     "Y"     2013-05-22    60
3 01-701-1294  5045 ACTOT         8 Week 8   WEEK 12       6 ""     ""      2013-06-14    83
4 01-701-1294  5045 ACTOT        16 Week 16  WEEK 12       6 "LOCF" "Y"     2013-06-14    83
5 01-701-1294  5060 ACTOT        24 Week 24  RETRIEVAL     9 ""     "Y"     2013-10-08   199
```

The same issue occurred for other subjects and resulted in the following discrepancies:

```
Not all Values Compared Equal
All rows are shown in table below

  ============================
   Variable  No of Differences
  ----------------------------
     AVAL           47
     CHG            47
     PCHG           47
    DTYPE           19
  ----------------------------
```

In the CDISC ADADAS, there are 19 subjects whose records have the incorrect DTYPE=LOCF value instead of the expected missing DTYPE, resulting in 47 records having incorrect AVAL/CHG/PCHG values for these subjects.

```
> diff <- diffdf(adadas, qc_adadas, keys = c("USUBJID", "PARAMCD", "AVISIT", "ADT"))
> count(diff$VarDiff_AVAL, USUBJID)
```

```
# A tibble: 19 × 2
   USUBJID         n
   <chr>       <int>
 1 01-701-1294     2
 2 01-701-1302     2
 3 01-703-1076     3
 4 01-704-1065     3
 5 01-704-1120     3
 6 01-705-1292     1
 7 01-705-1310     3
 8 01-708-1347     3
 9 01-709-1102     3
10 01-709-1259     2
11 01-710-1045     3
12 01-710-1278     3
13 01-710-1300     3
14 01-710-1315     2
15 01-714-1068     3
16 01-715-1107     2
17 01-716-1373     3
18 01-718-1172     2
19 01-718-1250     1
```

ADTTE

The R-generated ADTTE matches original ADTTE from CDISC pilot data except for minor SAS format discrepancies. Since this adtte was generated in R compared to SAS formats, the columns Type & Length in the define should be sufficient enough to describe the attributes of these variables.

```
> diffdf(adtte, qc_adtte, keys = c("STUDYID", "USUBJID", "PARAMCD", "SRCDOM", "STARTDT"))
Differences found between the objects!

A summary is given below.

There are columns in BASE and COMPARE with differing attributes !!
First 10 of 20 rows are shown in table below

  ==================================================
   VARIABLE  ATTR_NAME   VALUES.BASE  VALUES.COMP
  --------------------------------------------------
     AGE     format.sas     NULL           3
    AGEGR1   format.sas     NULL          $5
   AGEGR1N   format.sas     NULL           3
   EVNTDESC  format.sas     NULL          $25
    PARAM    format.sas     NULL          $32
   PARAMCD   format.sas     NULL          $4
     RACE    format.sas     NULL          $32
    RACEN    format.sas     NULL           3
    SAFFL    format.sas     NULL          $1
     SEX     format.sas     NULL          $1
  --------------------------------------------------
```

Label discrepancies

In pilot3, variable labels were updated per ADaM IG 1.1, which caused some discrepancies with original CDISC pilot data label.

Dataset	Variable	CDISC pilot data label	Pilot3 label
ADAE	ADURN	Analysis Duration (N)	AE Duration (N)
	ADURU	Analysis Duration Units	AE Duration Units
	AOCCFL	1st Occurrence of Any AE Flag	1st Occurrence within Subject Flag
ADADAS	ANL01FL	Analysis Record Flag 01	Analysis Flag 01
	ITTFL	Intent-to-Treat Population Flag	Intent-To-Treat Population Flag
ADTTE	SRCDOM	Source Data	Source Domain