Text File Comparison and Proc Compare

Steve Gibbs

- Validation is a big part of clinical trials data.
- One of the preferred methods is Independent Programming.
- The objective being to obtain a match between production and validation datasets.
- This doesn’t necessarily stop at dataset production but can also be used for Tables, Figures and Listings as well.

- The most common method in SAS to obtain validation/QC pass for a dataset is to obtain a match via Proc Compare.

- Typical desired output would be...

```
                          The SAS System

                     The COMPARE Procedure
           Comparison of WORK.BASE with WORK.COMPARE
              (Method=RELATIVE, Criterion=0.00001)

                        Data Set Summary

 Dataset                  Created           Modified  NVar     NObs

 WORK.BASE      29DEC05:11:00:29  29DEC05:11:00:29    34      377
 WORK.COMPARE   29DEC05:11:00:29  29DEC05:11:00:29    34      377


                        Variables Summary

             Number of Variables in Common: 34.


                       Observation Summary

                 Observation      Base  Compare

                 First Obs           1        1
                 Last  Obs         377      377

Number of Observations in Common: 377.
Total Number of Observations Read from WORK.BASE: 377.
Total Number of Observations Read from WORK.COMPARE: 377.

Number of Observations with Some Compared Variables Unequal: 0.
Number of Observations with All Compared Variables Equal: 377.

NOTE: No unequal values were found. All values compared are exactly equal.
```

- However this ideal doesn’t always materialize.
- In the event of which debugging is required to find out whether the problem is in the Production or QC program.
- A typical Proc Compare output in this case might be…

The COMPARE Procedure
Comparison of WORK.W1 with WORK.W3
(Method=EXACT)

All Variables Compared have Unequal Values

Variable	Type	Len	Ndif	MaxDif
v4	CHAR	20	91	
v5	CHAR	20	91	
v2	NUM	8	91	27.000
v3	NUM	8	91	21.000
v1	NUM	8	91	3.000

Value Comparison Results for Variables

Obs	Base Value v4	Compare Value v4
7	demo63	demo72
8	demo72	demo81
9	demo81	demo90
10	demo90	demo99
11	demo99	demo108
12	demo108	demo117
13	demo117	demo126
14	demo126	demo135
15	demo135	demo144
16	demo144	demo153
17	demo153	demo162
18	demo162	demo171
19	demo171	demo180
20	demo180	demo189
21	demo189	demo198
22	demo198	demo207
23	demo207	demo216
24	demo216	demo225
25	demo225	demo234
26	demo234	demo243
27	demo243	demo252
28	demo252	demo261

- Where Text File Comparison (TFC) can help is to show the actual data records which are disagreeing a little more clearly by stripping away a lot of the noise...

```
2      2 18 14 demo horizon
3      3 27 21 demo horizon
4      4 36 28 demo horizon
5      5 45 35 demo horizon
6      6 54 42 demo horizon
7  <!  7 63 49 demo horizon
8      8 72 56 demo horizon
9      9 81 63 demo horizon
10     10 90 70 demo horizon
11     11 99 77 demo horizon
12     12 108 84 demo horizon
13     13 117 91 demo horizon
14     14 126 98 demo horizon
15     15 135 105 demo horizon
16     16 144 112 demo horizon
17     17 153 119 demo horizon
18     18 162 126 demo horizon
19     19 171 133 demo horizon
20     20 180 140 demo horizon
21     21 189 147 demo horizon
22     22 198 154 demo horizon
23     23 207 161 demo horizon
24     24 216 168 demo horizon
25     25 225 175 demo horizon
26     26 234 182 demo horizon
27     27 243 189 demo horizon
28     28 252 196 demo horizon
29     29 261 203 demo horizon
30     30 270 210 demo horizon
31     31 279 217 demo horizon
32     32 288 224 demo horizon
33     33 297 231 demo horizon
34     34 306 238 demo horizon
35     35 315 245 demo horizon
36     36 324 252 demo horizon
37     37 333 259 demo horizon
38     38 342 266 demo horizon
39     39 351 273 demo horizon
40     40 360 280 demo horizon
41     41 369 287 demo horizon
42     42 378 294 demo horizon
43     43 387 301 demo horizon
44     44 396 308 demo horizon
45     45 405 315 demo horizon
46     46 414 322 demo horizon
47     47 423 329 demo horizon
48     48 432 336 demo horizon
49     49 441 343 demo horizon
50     50 450 350 demo horizon
```

```
1  1 9 7 demo horizon
2  2 18 14 demo horizon
3  3 27 21 demo horizon
4  4 36 28 demo horizon
5  5 45 35 demo horizon
6  6 54 42 demo horizon
7  8 72 56 demo horizon
8  9 81 63 demo horizon
9  10 90 70 demo horizon
10 11 99 77 demo horizon
11 12 108 84 demo horizon
12 13 117 91 demo horizon
13 14 126 98 demo horizon
14 15 135 105 demo horizon
15 16 144 112 demo horizon
16 17 153 119 demo horizon
17 18 162 126 demo horizon
18 19 171 133 demo horizon
19 20 180 140 demo horizon
20 21 189 147 demo horizon
21 22 198 154 demo horizon
22 23 207 161 demo horizon
23 24 216 168 demo horizon
24 25 225 175 demo horizon
25 26 234 182 demo horizon
26 27 243 189 demo horizon
27 28 252 196 demo horizon
28 29 261 203 demo horizon
29 30 270 210 demo horizon
30 31 279 217 demo horizon
31 32 288 224 demo horizon
32 33 297 231 demo horizon
33 34 306 238 demo horizon
34 35 315 245 demo horizon
35 36 324 252 demo horizon
36 37 333 259 demo horizon
37 38 342 266 demo horizon
38 39 351 273 demo horizon
39 40 360 280 demo horizon
40 41 369 287 demo horizon
41 42 378 294 demo horizon
42 43 387 301 demo horizon
43 44 396 308 demo horizon
44 45 405 315 demo horizon
45 46 414 322 demo horizon
46 47 423 329 demo horizon
47 48 432 336 demo horizon
48 49 441 343 demo horizon
49 50 450 350 demo horizon
```

- Proc Compare also provides summary information to help the user in this particular example:

NVar	NObs
5	100
5	97

- But if the number of observations are the same, it is still remarkably easy to pinpoint the differing records with TFC's relative dataset display.

- There are a variety of options available in Proc Compare to help with situations like this as well.
- One approach would be to use an ID variable
- This would take the previous output and transform it to:

```
                    Variables Summary

          Number of Variables in Common: 5.
          Number of ID Variables: 1.

           Comparison Results for Observations

Observation 7 in WORK.W1 not found in WORK.W3: v1=7.

Observation 32 in WORK.W1 not found in WORK.W3: v1=32.

Observation 94 in WORK.W1 not found in WORK.W3: v1=94.


                    Observation Summary

         Observation      Base  Compare  ID

         First Obs           1        1  v1=1
         Last  Obs         100       97  v1=100

Number of Observations in Common: 97.
Number of Observations in WORK.W1 but not in WORK.W3: 3.
Total Number of Observations Read from WORK.W1: 100.
Total Number of Observations Read from WORK.W3: 97.

Number of Observations with Some Compared Variables Unequal: 0.
Number of Observations with All Compared Variables Equal: 97.

NOTE: No unequal values were found. All values compared are exactly equal.
```

- Some ID variables are intuitive like Subject ID for example, others not so much e.g. covariates such as age group, biomarker type etc.

- In view of this the TFC approach can offer valuable exploratory information to home-in on these differing records.

- What type of code might you use to get these text outputs?

- An example would be:

```
%let outtfl=%str(C:\somwhere\C_077\DELIVERABLES);
data _null_;
  file "&outtfl\w1.txt";
  set w1;
  put v1 v2 v3 v4 v5;
run;

data _null_;
  file "&outtfl\w3.txt";
  set w3;
  put v1 v2 v3 v4 v5;
run;
```

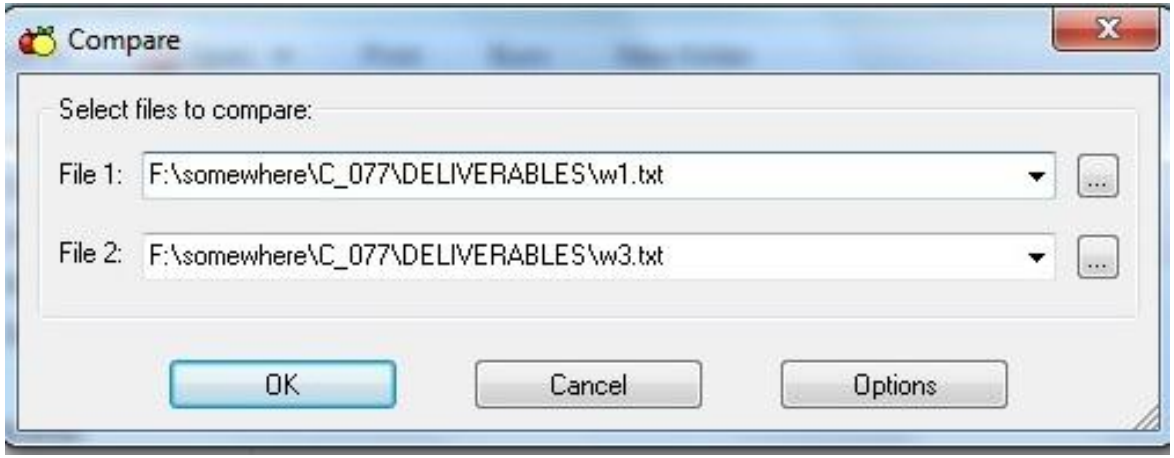
- Is the output pretty?
- Not always…

```
\clbrdrt\brdrs\brdrw15\brdrcf1\clbrdrb\brdrs\brdrw5\brdrcf1\clbrdrl\brdrs\brdrw15\brdrcf1\cltxlrtb\clvertalb\clcbpat17\cellx525
\clbrdrt\brdrs\brdrw15\brdrcf1\clbrdrb\brdrs\brdrw5\brdrcf1\clbrdrl\brdrs\brdrw5\brdrcf1\cltxlrtb\clvertalb\clcbpat17\cellx1190
\clbrdrt\brdrs\brdrw15\brdrcf1\clbrdrb\brdrs\brdrw5\brdrcf1\clbrdrl\brdrs\brdrw5\brdrcf1\cltxlrtb\clvertalb\clcbpat17\cellx2757
\clbrdrt\brdrs\brdrw15\brdrcf1\clbrdrb\brdrs\brdrw5\brdrcf1\clbrdrl\brdrs\brdrw5\brdrcf1\cltxlrtb\clvertalb\clcbpat17\cellx3312
\clbrdrt\brdrs\brdrw15\brdrcf1\clbrdrb\brdrs\brdrw5\brdrcf1\clbrdrl\brdrs\brdrw5\brdrcf1\cltxlrtb\clvertalb\clcbpat17\cellx3867
\clbrdrt\brdrs\brdrw15\brdrcf1\clbrdrb\brdrs\brdrw5\brdrcf1\clbrdrl\brdrs\brdrw5\brdrcf1\cltxlrtb\clvertalb\clcbpat17\cellx4885
\clbrdrt\brdrs\brdrw15\brdrcf1\clbrdrb\brdrs\brdrw5\brdrcf1\clbrdrl\brdrs\brdrw5\brdrcf1\cltxlrtb\clvertalb\clcbpat17\cellx6410
\clbrdrt\brdrs\brdrw15\brdrcf1\clbrdrb\brdrs\brdrw5\brdrcf1\clbrdrl\brdrs\brdrw5\brdrcf1\cltxlrtb\clvertalb\clcbpat17\cellx7100
\clbrdrt\brdrs\brdrw15\brdrcf1\clbrdrb\brdrs\brdrw5\brdrcf1\clbrdrl\brdrs\brdrw5\brdrcf1\cltxlrtb\clvertalb\clcbpat17\cellx7688
\clbrdrt\brdrs\brdrw15\brdrcf1\clbrdrb\brdrs\brdrw5\brdrcf1\clbrdrl\brdrs\brdrw5\brdrcf1\cltxlrtb\clvertalb\clcbpat17\cellx8378
\clbrdrt\brdrs\brdrw15\brdrcf1\clbrdrb\brdrs\brdrw5\brdrcf1\clbrdrl\brdrs\brdrw5\brdrcf1\cltxlrtb\clvertalb\clcbpat17\cellx8966
\clbrdrt\brdrs\brdrw15\brdrcf1\clbrdrb\brdrs\brdrw5\brdrcf1\clbrdrl\brdrs\brdrw5\brdrcf1\cltxlrtb\clvertalb\clcbpat17\cellx9656
\clbrdrt\brdrs\brdrw15\brdrcf1\clbrdrb\brdrs\brdrw5\brdrcf1\clbrdrl\brdrs\brdrw5\brdrcf1\cltxlrtb\clvertalb\clcbpat17\cellx10244
\clbrdrt\brdrs\brdrw15\brdrcf1\clbrdrb\brdrs\brdrw5\brdrcf1\clbrdrl\brdrs\brdrw5\brdrcf1\cltxlrtb\clvertalb\clcbpat17\cellx10934
\clbrdrt\brdrs\brdrw15\brdrcf1\clbrdrb\brdrs\brdrw5\brdrcf1\clbrdrl\brdrs\brdrw5\brdrcf1\clbrdrr\brdrs\brdrw15\brdrcf1\cltxlrtb\
```

- But computers seem to like it.

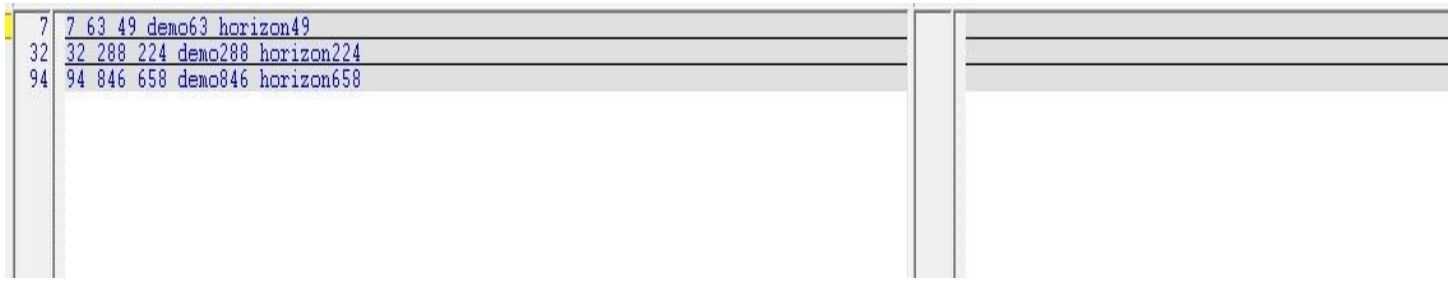
- What type of software might you use to compare them?
- Ultra Edit has the facility to do this.
- Another option would be Exam Diff (freeware)
- Both operate a similar procedure

- 1. Select your files to compare:



- 2. Click OK to compare and see output as illustrated earlier in slide 8.

- TFC Layout options:
- 1. Just show differences.



- 2. Just show matches.
- 3. Show everything.

- TFC Comparison options:
- 1. Ignore white space.
- 2. Ignore case.

- Why is TFC just a potential assistant to Proc Compare?
- The detail provided in Proc Compare is much more thorough looking at attributes such as variable type, length, format etc.
- Proc Compare is much more reliable and robust.
- Both Exam Diff and Ultra Edit can and do fail.
- An example of how they can fail is that the shading sometimes won't appear to denote differences when the server is busy, creating a false impression of a match.
- As a consequence they may most commonly be used to contribute to the comparison process and potentially save time in identifying where differences may exist prior to correction and final validation by Proc Compare.

REFERENCES:

Exam Diff
http://www.prestosoft.com/edp_examdiff.asp

Ultra Edit
http://www.ultraedit.com

- Questions?