

Validation of Clinical Output: More Quality with Less Effort

Jules van der Zalm, OCS Biometric Support, the Netherlands

ABSTRACT

When validating clinical data and programs, most clinical programmers manually check their data or manually compare the actual results to their validation output. This is a time consuming and error prone process whilst the right use of the right tools can result in a significant reduction of effort and an improvement in quality.

In this paper you will be shown how commonly available tools such as advanced text editors, text compare programs and spreadsheets can be of use when validating clinical output. Both the paper and the presentation will discuss ways to use these tools by describing under which circumstances they can be best applied.

INTRODUCTION

We all come across it, and it may just be one of the most time-consuming activities in the creation of reports and listings in pharmaceutical companies: validation, or quality assurance if you like. In this paper, we'll call it validation. Validation is performed from the early to the late stages of the programming of clinical trials and it is often an annoyance as it takes up so much time in the process. It is not only the process of, for example, comparing final output to the actual source data, but is also performed continuously while the applications are being programmed. The latter we'll call pre-validation and that's what this paper is about.

Pre-validation is often a manual process which can be very time consuming, but it can on the other hand take away many problems of the final validation process. There are many tools and applications available that make the pre-validation process more efficient and more effective. In this paper we'll use UltraEdit¹, CSDiff² and Microsoft Excel.

UltraEdit is a text-editor like Notepad, but it is more extensive with useful functionality like code highlighting, column mode selection and editing, macros, extensive find- and replace, code templates, formatting options and much more. CSDiff is a tool that can be used to compare different versions of text files, in this case reports and listings. When opened, it shows the base revision of a file and highlights the changed, removed or added elements as compared to the compared revision. Excel, the world's best known spreadsheet application, probably doesn't need any introduction. Please bear in mind that this paper was written for Windows users. Not all applications used may be available on other operating systems.

Important note: all data sets, tables and listings used in this paper and its accompanying presentation are based on dummy data. Data may be imprecise, incorrect or unrealistic. They've been created solely to support the methods set out in the paper.

PRE-VALIDATION

Pre-validation is the continuous process that takes place during the programming of (SAS-) programs in which result data, tables or listings are compared to the source data or where the results are read and recalculated in a different way to verify their correctness. It's mostly undocumented and not considered as an official validation or approval of the programs.

The following paragraphs will illustrate a couple of situations in which the previously mentioned applications take away parts of the hassle involved in pre-validation.

¹ UltraEdit, by IDM Computer Solutions, Inc. (www.ultraedit.com, shareware version available)

² CSDiff, by ComponentSoftware Inc. (www.componentsoftware.com/Products/CSDiff, freeware)

PhUSE 2008

COUNTING SUBJECTS IN A LISTING

A type of validation that is often performed on data listings is to count the number of subjects in a listing. Some listings are required to contain only information of a subset of subjects instead of all of them, like the Per-Protocol group or the All-Subjects-Randomized group. Other listings are, for example, restricted to subjects that had a serious adverse event or reached a certain stage of the study.

To make sure that all required subjects are in the listing, it's necessary to count the number of subjects in the listing to compare it to the expected number of subjects, or to obtain a list of all subjects that are in the listing.

Let's assume we have the following listing:

		10	20	30	40	50	60	70
01	----- ----- ----- ----- ----- ----- ----- -----							
02	Listing 14.1.1	Vital signs at screening.						
03		Per-Protocol Group.						
04								
05				Heart	Systolic	Diastolic		
06	Subject	Actual	Planned	rate	blood	blood		
07	ID	treatment	assess.	(BPM)	pressure	pressure	Height	Weight
08					(mmHg)	(mmHg)	(CM)	(KG)
09								
10	101001	Placebo	Pre-inj	99	68	120	184	73.9
11		Placebo	14 Days	96	69	118	184	73.3
12								
13	101002	Placebo	Pre-inj	79	70	116	155	41.6
14		Placebo	14 Days	79	69	114	155	39.8
15								
16	101003	Placebo	Pre-inj	96	86	121	181	70.7
17		Placebo	14 Days	99	85	120	181	73.6
	... and so on.							

Of course, such vital signs listings would run over many pages, especially in larger studies with many subjects and many assessments. Manually counting all unique subjects would take a lot of time and could even be inaccurate, because mistakes are easily made.

With UltraEdit, it is very easy to count the number of subjects in a listing, regardless of the layout of the listing's titles, headers and footnotes. UltraEdit's *sort* functionality allows you to sort on selected 'columns' (as opposed to the complete lines) in the file. So, when the ASCII output is opened in UltraEdit and sorted by the columns that the subject number is in (in the example, that would be columns 3 through 8 as seen in the ruler above the listing), the result will look like the extract below.

Note: when there are multiple values with the same value in the chosen range of columns, UltraEdit does not sort those records in a specific way. For cases where the subject number appears more than once in a listing, UltraEdit has the functionality to remove duplicates, like the NODUPKEY option in PROC SORT.

PhUSE 2008

```

01          10          20          30          40          50          60          70
02  -----|-----|-----|-----|-----|-----|-----|-----
03          Per-Protocol Group.
04          Heart blood blood
05          Placebo 14 Days 79 69 114 155 39.8
06          Placebo 14 Days 96 69 118 184 73.3
07          Systolic Diastolic
08
09
10  101001 Placebo Pre-inj 99 68 120 184 73.9
11  101002 Placebo Pre-inj 79 70 116 155 41.6
12  101003 Placebo Pre-inj 96 86 121 181 70.7
13  ID treatment assess. (BPM) (mmHg) (mmHg) (CM) (KG)
14  Subject Actual Planned rate pressure pressure Height Weight
15
16 Listing 14.1.1 Vital signs at screening.
17          Placebo 14 Days 99 85 120 181 73.6

```

By removing all the lines that do not contain a subject number, the output can be easily filtered to eliminate unnecessary data. In this example, the lines that would remain are lines 10, 11 and 12. The remainder would then consist of these three lines, and the line number in the left margin would tell you how many subjects there are in the listing.

In this small example it would of course be easy to count the number of subjects manually, but imagine a similar listing with hundreds or thousands of subjects. It would take up really much time to count them all.

UltraEdit performs really well on this, even with larger listings. For example, sorting a 70.000 lines (nearly 1500 pages) file with more than 14.000 unique subjects takes no more than 7 seconds on a modern computer³.

³ The computer used in this benchmark is an AMD Athlon XP 3200+ with 2 GB of RAM running Windows XP SP3. The version of UltraEdit is 14.00b.

PhUSE 2008

CALCULATING PERCENTAGES

Many tables display numbers in the form of percentages, like tables on Adverse Events or Concomitant Medication, but also in less standard or study specific tables. In most cases, these percentages should add up to 100%. Since such tables can become fairly large, counting these percentages manually would take up large amounts of time.

In a situation as described above, Microsoft Excel can come in very handy. Let's assume that you have created a table that looks like this.

		10	20	30	40	50	60	70
01	Table 14.3.1	Number of subjects with at least one (serious)						
02		adverse event.						
03		Per-Protocol Group.						
04								
05	+	+-----+-----+-----+-----+-----+-----+						
06						Fic-Treat		Placebo
07						+-----+-----+		
08						(N=691)		(N=700)
09						+-----+-----+		
10						N	%	N
11	+	+-----+-----+-----+-----+-----+-----+						
12		No adverse event				550	39.5	571
13		Non-serious adverse event				124	8.9	114
14		Serious adverse event				16	1.2	13
15		Serious adverse event leading to death				1	0.1	2
16	+	+-----+-----+-----+-----+-----+-----+						

In this small example manually adding up the percentages would not really be a big issue, but it's a good means of demonstrating the method. The method itself is based on Excel's functionality to quickly look up values like sum and average in the status bar. The status bar is the slim horizontal bar below the scroll bar and the tab selection.

The method starts with UltraEdit. Using UltraEdit's '*Column mode*' (accessible through the icon bar or by pressing the key combination Alt+C) you can easily select and copy the involved values without touching the parts of the table you don't need. Once you've selected the values – one block at a time - paste the values to an arbitrary place on the spreadsheet. After you've copied all blocks of values (two in this example), select all the values on the spreadsheet. Excel now shows you the sum (by default) of the selected number and thus shows you the sum of the selected percentages, which should be 100 (or 99.9 as SAS results do not always add up to 100 due to rounding of the values).

This same method can also be applied to count the number of subjects in a table. Instead of selecting the percentages, you now select the values under the column 'N' and copy them to Excel.

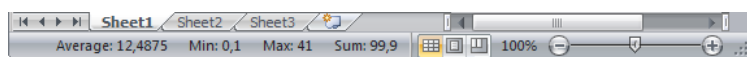
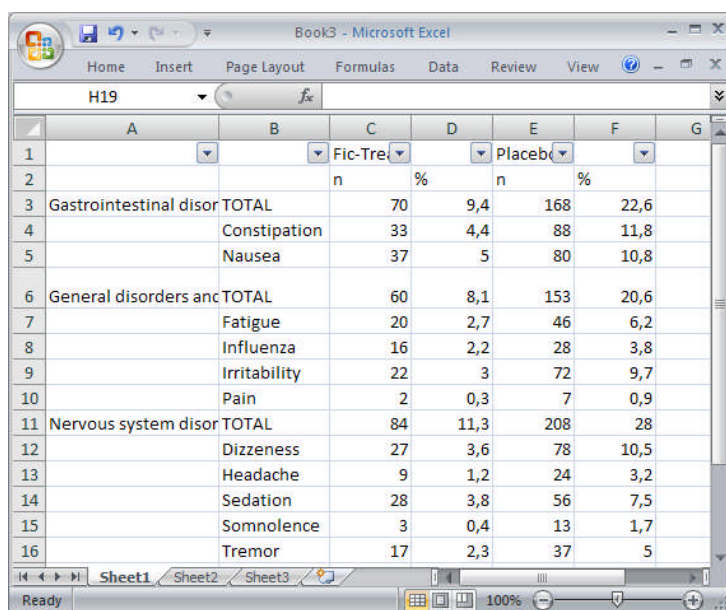


Image 1: The status bar in Excel 2007 showing the average, minimum, maximum and sum. The scroll bar and tab selection are visible above the status bar.

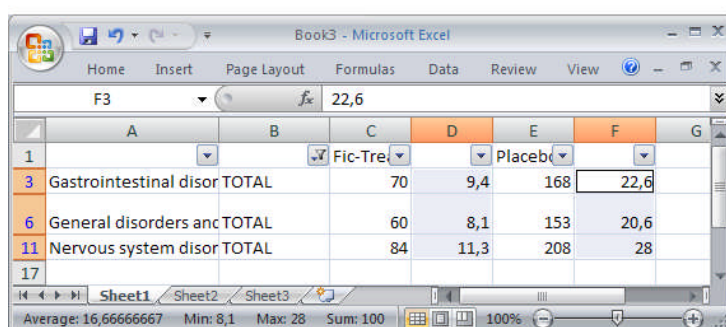
In versions before Excel 2007 you can right-click on the value in the status bar to choose to display a different type of value (e.g. sum, average, minimum, maximum). Excel 2007 allows you to display all desired value types at once (see *Image 1*) so requires you to set your preference only once.

Instead of using the ASCII files, this functionality also works really well on RTF output. In fact, that's even easier. The table functionality in Word is fully compatible with the cells layout of a spreadsheet in Excel, so when you copy and paste a complete table from Word (RTF) to Excel, you can use all the Excel functionality to perform calculations.

PhUSE 2008



	A	B	C	D	E	F
			Fic-Tre		Placebo	
			n	%	n	%
3	Gastrointestinal disorder	TOTAL	70	9,4	168	22,6
4		Constipation	33	4,4	88	11,8
5		Nausea	37	5	80	10,8
6	General disorders and	TOTAL	60	8,1	153	20,6
7		Fatigue	20	2,7	46	6,2
8		Influenza	16	2,2	28	3,8
9		Irritability	22	3	72	9,7
10		Pain	2	0,3	7	0,9
11	Nervous system disorder	TOTAL	84	11,3	208	28
12		Dizziness	27	3,6	78	10,5
13		Headache	9	1,2	24	3,2
14		Sedation	28	3,8	56	7,5
15		Somnolence	3	0,4	13	1,7
16		Tremor	17	2,3	37	5



	A	B	C	D	E	F
			Fic-Tre		Placebo	
			n	%	n	%
3	Gastrointestinal disorder	TOTAL	70	9,4	168	22,6
6	General disorders and	TOTAL	60	8,1	153	20,6
11	Nervous system disorder	TOTAL	84	11,3	208	28

Average: 16,6666667 Min: 8,1 Max: 28 Sum: 100

Image 2: Using the Filter-function in Microsoft Excel to only show the 'TOTAL'-groups in an Adverse Events RTF table.

In the above example (Image 2), an RTF table is copied from Microsoft Word to Microsoft Excel. The upper screen shows the complete adverse events incidence table. Column A shows the High Level Group Term, column B shows the Preferred Term. Columns C and E show the number of subjects with that adverse event for the respective treatment groups, columns D and F show the percentages.

The filter functionality (simply called 'Filter' in Excel 2007, 'AutoFilter' in earlier versions) allows you to show only a selected subset of records in the Excel spreadsheet. The functionality has been enabled on the spreadsheet, which can be identified by the downward arrows in the cells of Row 1. When such an arrow is clicked, a menu appears containing all distinct values in that column. When a value is selected, only those rows which contain that value in the column are shown. In the example, the value 'TOTAL' is selected which results in the spreadsheet as seen in the lower image. When the percentages in column D and F are then selected, only the percentages of the TOTAL-groups are used in the calculations of the average, minimum, maximum and sum. In the example, the status bar shows all these derivatives.

Please bear in mind that there's a difference in the interpretation of numbers between countries. Some countries require a comma as decimal separator, other require a dot. If the SAS table you're using uses the wrong separator, use the UltraEdit 'Replace' functionality to replace the dots with commas, or vice versa, before copy-and-pasting them to Excel, or use Excel's 'Replace'-functionality afterwards. What separator your computer requires depends on the settings in Microsoft Windows. When the wrong separator is used, Excel interprets the values as text and cannot perform the calculations.

PhUSE 2008

COMPARING FILES

Programs change often. When they're in development, different copies can roam around in your development environment. Or when you're recycling your programs over (almost) identical studies, several different versions can exist in the study folders. It's hard keeping track of all these versions, which one is the newest and what changes have been made as compared to the version you're currently looking at.

A handy utility in such situations is Component Software's Diff, better known as CSDiff. CSDiff compares two text files and points out all the changed, removed and added parts in the 'compare revision' as compared to the 'base revision'. An example is shown in Image 3.

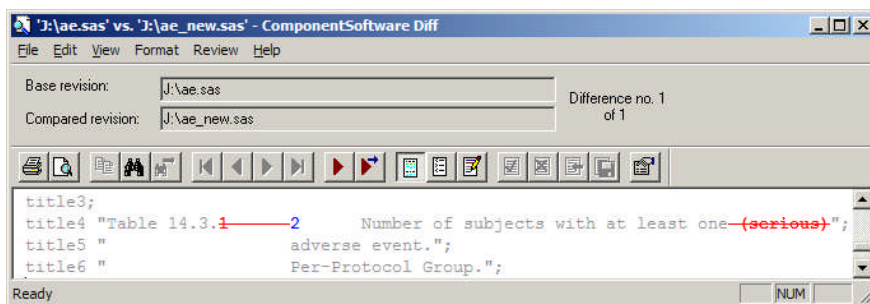


Image 3: Comparing two versions of an Adverse Events listing program.

The example shows the comparison of two versions of an Adverse Events listing program. The base revision is file 'ae.sas', the compared revision if 'ae_new.sas'. In the line starting with 'Title 4', the red struck through words are the ones that are deleted from the base revision, the blue bold words are those added in the compared revision. So, in the example, the table number was changed from '14.3.1' to '14.3.2' and the text '(serious)' was removed from the title.

UltraEdit-32 has comparable functionality with the 'Compare Files' functionality in the 'File'-menu, although this is only working well since version 13.00 and still not as powerful as CSDiff's functionalities, as it's only able to compare files on line level, not on character level.

PhUSE 2008

CONCLUSION

The time required to perform pre-validation while programming your clinical trials in SAS can be significantly reduced, just by choosing the right tools and applications and using them correctly. With some creativity the standard functionality of some applications can become a really great help in the everyday routines of validation.

Be creative!

ACKNOWLEDGEMENTS

I would like to thank the following people for inspiring me and for reviewing this paper and providing feedback.

Raymond Ebben, OCS Consulting

Bas van Bakel, OCS Consulting

CONTACT INFORMATION

Your questions and comments are valued and encouraged. Contact the author at:

Author name	:	Jules van der Zalm
Company	:	OCS Consulting
Address	:	P.O. Box 490 5240 AL Rosmalen The Netherlands
Work phone	:	+31 (0)73 523 6000
Facsimile	:	+31 (0)73 523 6600
E-mail	:	sasquestions@ocs-consulting.com
Website	:	www.ocs-consulting.nl

Brand and product names are trademarks of their respective companies.