PHUSE/FDA CSS 2021 PP023

BEACONCURE

New Uses for Automation in the **Output Validation Process**

Ilan Carmeli, Hugh Donovan, Beaconcure

The tables that appear in the deliverable are the result of a four-step process:

1. Specification of the List of Tables (LoT) which defines the tables to be produced
2. Definition of the Tables, Listings and Figures Shells (TLFs) as a part of the Statistical Analysis Plan or a separate document
3. Generation of the tables
4. Generation of the in-text tables

Each step can be checked using automated checks incorporating machine learning and artificial intelligence. Similar techniques can be used to check for consistency of the deliverables from each step, e.g. are the in-text tables discrepant from the end-of-text tables, are the table shells the same as the tables, and have all tables in the LoT been produced with the same headings.

LoT

VERIFY BY BEACONCURE

TLFs

MOCK SHELLS

VERIFY BY BEACONCURE

TLFs

The current gold standard for validation in the industry is 100% double programming of all analysis datasets, tables, listings and figures. In a survey conducted by Beaconcure all 25 respondents use double programming for at least some of the outputs.
This approach, however, does not resolve cross-table discrepancies, and typically does not account for unanticipated data changes during clinical trials. Also, as it is a manual exercise, it is also subject to human error.

Today, with the support of ML-driven technology, the industry would do well to consider automating the output validation process.
An automation solution developed by Beaconcure, checks outputs in exactly the same way as figures within tables are commonly compared today. This technology can be used multiple times as data accumulates, identifying discrepancies in the output. One example of within table checks and discrepancies are shown below:

WITHIN TABLE VALIDATION

Table 6
Drug Protocol Demo
Shift Table of Lab Data - White Blood Cell Count at Last Visit
Drug A (N=118)

		Missing	Below Normal	Within Normal	Above Normal	Total
Baseline	Missing	1 (0.8%)	0 (0.0%)	2 (1.7%)	0 (0.0%)	3 (2.5%)
WBC	Below Normal Range	0 (0.0%)	12 (10.1%)	0 (0.0%)	0 (0.0%)	12 (10.1%)
	Within Normal Range	0 (0.0%)	3 (2.5%)	63 (52.9%)	6 (5.0%)	72 (60.5%)
	Above Normal Range	0 (0.0%)	0 (0.0%)	12 (10.1%)	20 (16.8%)	32 (26.9%)
	Total	1 (0.8%)	15 (12.6%)	77 (64.7%)	26 (21.8%)	118 (99.2%)

FOOTNOTES
Creation: 3FEB2020 (14:50) Source Data: abce Date of Generation: 13FEB2020 (18:37)

Despite its limitations, double programming is used for individual tables, but it does not compare the output across tables. Within an individual table, the two programs may produce the same answer, but it may not be correct. This is something we have seen in our analysis of tables. To overcome this limitation of double programming, statisticians perform a manual check for consistency across tables. While this will improve quality, there are shortcomings.

Verify, in addition to performing within table checks, can perform cross-table checks quickly and consistently for all deliverables. This is achieved by running a set of standard cross table checks.

CROSS TABLE CHECK

The discrepancies are highlighted in the output so that the reviewer can immediately see where the discrepancies are.

Table 14.1.1.2
Subject Disposition

	Drug A	Drug B	Total
Randomized	119	117	236
Received at least one dose	118	118	236
Completed 4 weeks	89	87	176
Completed 8 weeks	82	75	157
Completed Study	75	69	144
Per Protocol	48	39	87

Note: Subject 003-005 was randomized to Drug A but received Drug B in error.
Cutoff date: 03MAY2021 Snapshot date: 03MAY2021

Table 14.1.1.4
Demographics (Received at Least One Dose)

		Drug A	Drug B	Total
Gender				
	Male	60 (51.3%)	60 (50.8%)	120 (50.8%)
	Female	58 (48.7%)	58 (49.2%)	116 (49.2%)
Age				
	<18	1 (0.8%)	0	1 (0.4%)
	18-30	34 (28.6%)	33 (28.2%)	68 (28.8%)
	31-45	36 (30.3%)	37 (31.6%)	73 (30.9%)
	46-65	46 (38.7%)	47 (40.2%)	93 (39.4%)
	>65	1 (0.8%)	0	1 (0.4%)
Weight				
	n	118	115	233
	Mean	75.12	75.34	77.79
	SD	7.34	7.52	7.74
	Median	77.8	78.1	77.9
	Min	62	67	62
	Max	88	89	89

Note: Subject 003-005 was randomized to Drug A but received Drug B in error.
Cutoff date: 03MAY2021 Snapshot date: 03MAY2021

Table 14.1.2.4
Overall Summary of Treatment Emergent Adverse Events (Safety Population)

	Drug A (N=119)	Drug B (N=117)	Total (N=236)
Any AEs	73 (61.3%)	88 (75.2%)	161 (68.2%)
0-4 Weeks:			
Any TEAEs	75 (63.0%)	85 (72.6%)	156 (66.1%)
Serious TEAEs	3 (2.5%)	17 (14.3%)	7 (5.9%)
Drug-Related TEAEs	17 (14.3%)	27 (23.1%)	44 (18.6%)
Discontinued due to TEAEs	7 (5.9%)	9 (7.7%)	16 (6.8%)
4-8 Weeks:			
Any TEAEs	48 (40.3%)	38 (32.5%)	86 (36.4%)
Serious TEAEs	1 (0.8%)	3 (2.6%)	4 (1.7%)
Drug-Related TEAEs	12 (10.1%)	20 (17.1%)	40 (16.9%)
Discontinued due to TEAEs	3 (2.5%)	4 (3.4%)	7 (3.0%)
8-12 Weeks:			
Any TEAEs	25 (21.0%)	20 (17.1%)	45 (19.1%)
Serious TEAEs	0 (0.0%)	1 (0.9%)	1 (0.4%)
Drug-Related TEAEs	7 (5.9%)	5 (4.3%)	12 (5.1%)
Discontinued due to TEAEs	1 (0.8%)	1 (0.9%)	2 (0.8%)
0-12 Weeks:			
Any TEAEs	83 (69.7%)	93 (79.5%)	176 (74.6%)
Serious TEAEs	3 (2.5%)	7 (6.0%)	10 (4.2%)
Drug-Related TEAEs	28 (23.5%)	30 (25.6%)	58 (24.6%)
Discontinued due to TEAEs	11 (9.2%)	14 (12.0%)	25 (10.6%)

Note: Subject 003-005 was randomized to Drug A but received Drug B in error.
The Safety Population is all subject who received at least one does of medication.
Any AE is any adverse event that began after the beginning of the screening period.
Any TEAE is any adverse event that began after the first dose.
Cutoff date: 03MAY2021 Snapshot date: 03MAY2021

LoT

Specification of the List of Tables (LoT) which defines the tables to be produced is compared to identify discrepant titles as shown below:

Table Number	Table Title
14.1.1.4	Demographics (safety population)

Table 14.1.1.4
Demographics (Received at least one dose)

		Drug A (N=119)	Drug B (N=117)	Total (N=236)
Gender	Male	60 (50.4%)	60 (51.3%)	120 (50.8%)
	Female	59 (49.6%)	57 (48.7%)	116 (49.2%)
Age	<18	1 (0.8%)	0 (0.0%)	1 (0.4%)
	18-30	34 (28.6%)	33 (28.2%)	68 (28.8%)
	31-45	36 (30.3%)	37 (31.6%)	73 (30.9%)
	46-65	46 (38.7%)	47 (40.2%)	93 (39.4%)
	>65	1 (0.8%)	0 (0.0%)	1 (0.4%)
Weight (kg)	n	118	115	233
	Mean	75.12	75.34	77.79
	SD	7.34	7.52	7.44
	Median	77.8	78.1	77.9
	Min	62	67	60
	Max	88	89	89

FOOTNOTES
Creation: 3FEB2020 (16:40) Source Data: abcd Date of Generation: 13FEB2020 (18:37)

CONCLUTION

The algorithms determine whether a certain check was relevant for particular tables, so there is no need to do study specific configuration. If a new check is required, it can be added to the standard set so it is available for all studies, not just one. This automated approach requires a statistician to resolve discrepancies raised by Verify, a task often requiring judgment and knowledge of the study. There would also still be a benefit to a statistician reviewing the main results of the study at a high level, for example are the results in the tables consistent with the p-values. Automation, while not replacing the statistician can:

- Increase the quality of statistical analysis
- Check the output comprehensively and consistently, both within and across tables
- Greatly reduce the time and effort to perform checks
- Improve the productivity of statisticians and improve their job satisfaction
- Eliminate all manual steps in the output validation process workflow