PharmaSUG 2022 - Paper AI- 010

Data Mining of Tables: The Barrier for Automation

Ilan Carmeli, Co-Founder and VP Product at Beaconcure

ABSTRACT

The outputs of statistical analyses are static and considered to be unstructured/semi-structured files. Verify by Beaconcure was developed with a unique data processing pipeline in which clinical data tables are mined not only for cell values, but also for clinical context. The objective of the following paper is to provide an overview of current challenges working with unstructured/semi-structured formats, and introduce readers to Beaconcure's solution for table mining.

INTRODUCTION

The statistical analysis outputs of clinical trials data are often validated manually by SAS programmers and biostatisticians. The first reason for that is that the process of generating and validating statistical analyses outputs is not standard and could not be easily programmed. Specifications and definitions are written on a case-by-case-basis, and shared amongst the programmers in non-standard documents, emails, and spreadsheets. Secondly, the output of statistical analyses are considered to be unstructured/semi-structured files (PDF, RTF, HTML, etc). Metadata (information about the data) is missing, and information regarding the data hierarchy within and between the outputs does not exist. Figure 1 shows the results from an industry survey Beaconcure conducted, where pharmaceutical companies and Contract Research Organizations (CROs) were asked about their most commonly used output file formats.

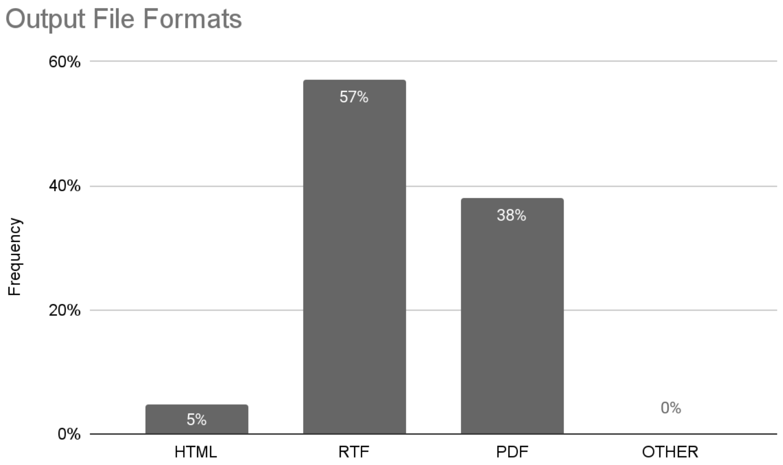


Figure 1: "Which output file formats are most commonly used in the industry?" Results from an industry survey (2022).

Additional challenges that exist with using unstructured/semi-structured files are lack of proper revision history, missing consolidated shared feedback, and lack of focal area of communication between CROs and the sponsors.

In order to develop any type of an automated QC tool, an automated solution for converting static files into machine readable format needs to be developed. The steps to follow are:

1. Parsing: converting raw files into an abstract structure
2. Utilizing information contained in the files, such as abbreviations, clinical terms, synonyms, etc.
3. Classification of headers identifies column and row headers automatically
4. Cells characteristics: adding the metadata information to each cell

As a result, the information of every cell resides in a structured database, which can be easily retrieved and used for various purposes.

Verify was developed with such mining capabilities; Verify accepts static data structures (statistical data outputs) and is able to automatically parse and identify the hierarchy of cells. For example, table 1 is a static table uploaded to Verify in RTF.

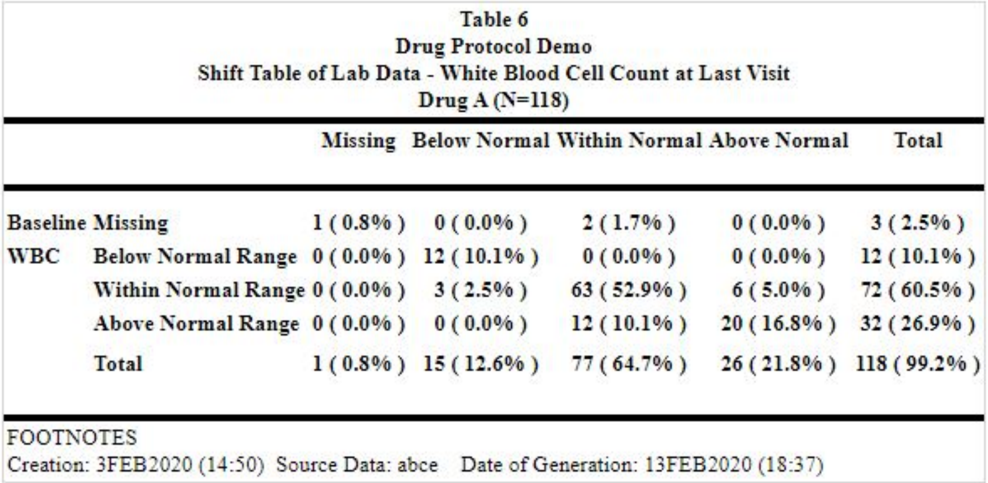


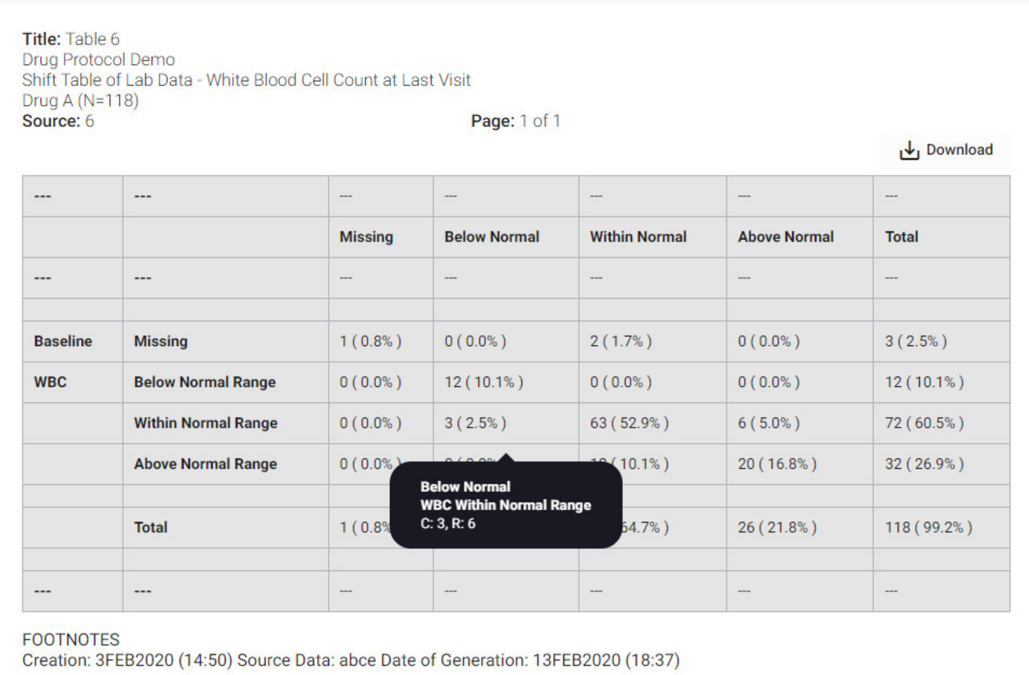
Table 6
Drug Protocol Demo
Shift Table of Lab Data - White Blood Cell Count at Last Visit
Drug A (N=118)

		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)

Table 1: static output of clinical trial data in HTML format.

The post-table mining process done by Verify outputs table 2, where cells have contextual information in addition to cell values.



Title: Table 6
Drug Protocol Demo
Shift Table of Lab Data - White Blood Cell Count at Last Visit
Drug A (N=118)
Source: 6 **Page:** 1 of 1

---	---	---	---	---	---	---
		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%)	[illegible]	[illegible] (10.1%)	20 (16.8%)	32 (26.9%)
	Total	1 (0.8%	[illegible]	[illegible]4.7%)	26 (21.8%)	118 (99.2%)
---	---	---	---	---	---	---

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

Table 2: Verify mining process of a static table.

THE CHALLENGE IN CREATING TABLES IN MACHINE-READABLE FORMAT

Machine-readable data is a data format that can be easily read and processed (identified and extracted) by a computer without human intervention, including text and their internal structure, while ensuring no semantic meaning is lost. Machine-readable data must be structured data. Presenting information in tables is very challenging for automatic processing, since the traditional text mining techniques fail to capture the meaning of the data. Clinical data tables are usually unstructured files, designed for the human eye. For the machine to understand the numbers and values we will have to use advanced techniques of table mining. The table below (Table 3) is an example of a typical clinical data table. It is clear for humans that the value ‘15’ highlighted twice in the table has a different meaning in each cell, however, a computer cannot distinguish between them. The next section will go over the details of Beaconcure’s table mining process.

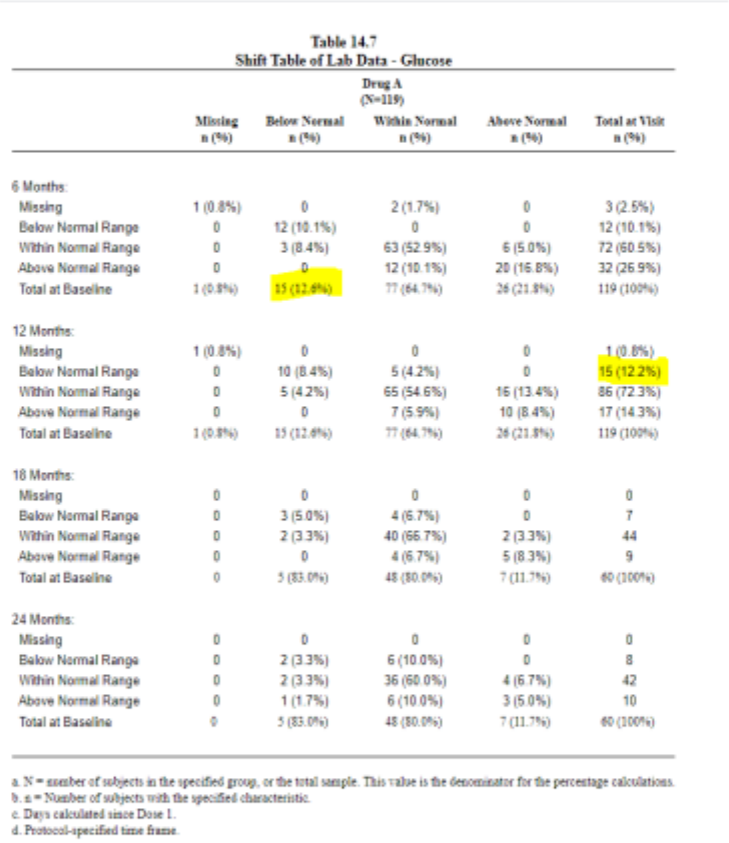


Table 14.7
Shift Table of Lab Data - Glucose

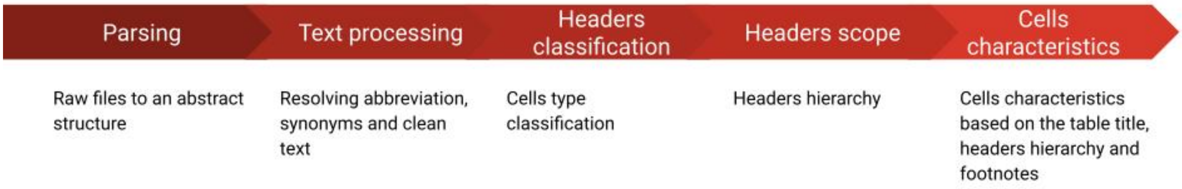
	Drug A (N=119)				
	Missing n (%)	Below Normal n (%)	Within Normal n (%)	Above Normal n (%)	Total at Visit n (%)
6 Months:					
Missing	1 (0.8%)	0	2 (1.7%)	0	3 (2.5%)
Below Normal Range	0	12 (10.1%)	0	0	12 (10.1%)
Within Normal Range	0	3 (8.4%)	63 (52.9%)	6 (5.0%)	72 (60.5%)
Above Normal Range	0	0	12 (10.1%)	20 (16.8%)	32 (26.9%)
Total at Baseline	1 (0.8%)	15 (12.6%)	77 (64.7%)	26 (21.8%)	119 (100%)
12 Months:					
Missing	1 (0.8%)	0	0	0	1 (0.8%)
Below Normal Range	0	10 (8.4%)	5 (4.2%)	0	15 (12.2%)
Within Normal Range	0	5 (4.2%)	65 (54.6%)	16 (13.4%)	86 (72.3%)
Above Normal Range	0	0	7 (5.9%)	10 (8.4%)	17 (14.3%)
Total at Baseline	1 (0.8%)	15 (12.6%)	77 (64.7%)	26 (21.8%)	119 (100%)
18 Months:					
Missing	0	0	0	0	0
Below Normal Range	0	3 (5.0%)	4 (6.7%)	0	7
Within Normal Range	0	2 (3.3%)	40 (66.7%)	2 (3.3%)	44
Above Normal Range	0	0	4 (6.7%)	5 (8.3%)	9
Total at Baseline	0	5 (83.0%)	48 (80.0%)	7 (11.7%)	60 (100%)
24 Months:					
Missing	0	0	0	0	0
Below Normal Range	0	2 (3.3%)	6 (10.0%)	0	8
Within Normal Range	0	2 (3.3%)	36 (60.0%)	4 (6.7%)	42
Above Normal Range	0	1 (1.7%)	6 (10.0%)	3 (5.0%)	10
Total at Baseline	0	5 (83.0%)	48 (80.0%)	7 (11.7%)	60 (100%)

a. N = number of subjects in the specified group, or the total sample. This value is the denominator for the percentage calculations.
b. n = Number of subjects with the specified characteristic.
c. Days calculated since Dose 1.
d. Protocol-specified time frame.

Table 3: An unstructured table with two cells highlighted with equivalent cell values (15) but different contextual meanings.

THE SOLUTION - TABLE PROCESS PIPELINE

Clinical data tables exist in different file and styling formats. Beaconcure's table mining standardizes the data to an abstract hierarchy repository. Figure 2 below shows the table process pipeline in which raw unstructured files are converted to structured, machine-readable data. Our unique solution includes Machine Learning (ML) models trained on clinical data tables for cell classification (header cells vs. data cells) and data hierarchy.

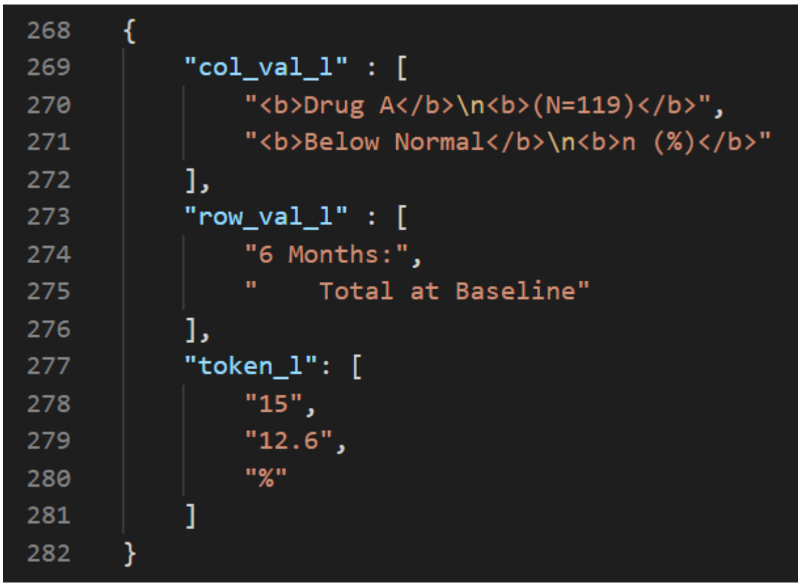


Parsing
Text processing
Headers classification
Headers scope
Cells characteristics
Raw files to an abstract structure
Resolving abbreviation, synonyms and clean text
Cells type classification
Headers hierarchy
Cells characteristics based on the table title, headers hierarchy and footnotes

Figure 2: Table process pipeline

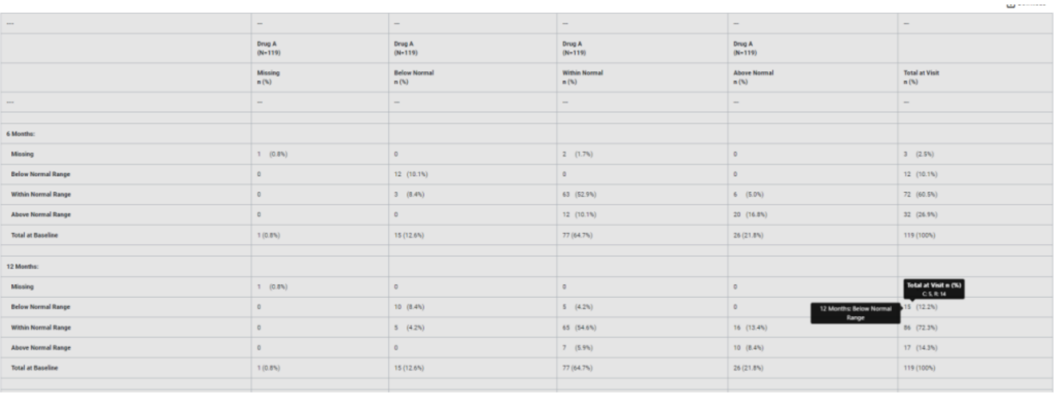
DATA REPRESENTATION

Table mining via the Beaconcure 'Table Process Pipeline' enables the machine to understand the data within the tables. Going back to the problem mentioned earlier, in which the value '15' appears twice in the same table and has a different meaning in the different cells. A human is able to immediately understand the context of each cell based on its referenced headers, while a computer cannot distinguish between the contextual value of two identical cell values. In Beaconcure's database, the representation of the data cells includes additional attributes; the corresponding column headers and row headers relate to the values that give the meaning to its data tokens. Figures 3 and 4 show a snapshot of the data representation for one of the cells that have the value '15' in the database and Beaconcure's analysis view, respectively.



```
{
    "col_val_1" : [
        "<b>Drug A</b>\n<b>(N=119)</b>",
        "<b>Below Normal</b>\n<b>n (%)</b>"
    ],
    "row_val_1" : [
        "6 Months:",
        "     Total at Baseline"
    ],
    "token_1": [
        "15",
        "12.6",
        "%"
    ]
}
```

Figure 3: Data representation of a cell in the database post-processing.



---	—	—	—	—	—
	Drug A (N=119)	Drug A (N=119)	Drug A (N=119)	Drug A (N=119)	
	Missing n (%)	Below Normal n (%)	Within Normal n (%)	Above Normal n (%)	Total at Visit n (%)
---	—	—	—	—	—
6 Months:					
Missing	1 (0.8%)	0	2 (1.7%)	0	3 (2.5%)
Below Normal Range	0	12 (10.1%)	0	0	12 (10.1%)
Within Normal Range	0	3 (8.4%)	63 (52.9%)	6 (5.0%)	72 (60.5%)
Above Normal Range	0	0	12 (10.1%)	20 (16.8%)	32 (26.9%)
Total at Baseline	1 (0.8%)	15 (12.6%)	77 (64.7%)	26 (21.8%)	119 (100%)
12 Months:					
Missing	1 (0.8%)	0	0	0	
Below Normal Range	0	10 (8.4%)	5 (4.2%)	0	15 (12.2%)
Within Normal Range	0	5 (4.2%)	65 (54.6%)	16 (13.4%)	86 (72.3%)
Above Normal Range	0	0	7 (5.9%)	10 (8.4%)	17 (14.3%)
Total at Baseline	1 (0.8%)	15 (12.6%)	77 (64.7%)	26 (21.8%)	119 (100%)

Figure 4: Beaconcure analysis view of the cell value '15' post-processing.

DATA VISUALIZATION

Table 4 shows an example of how tables are visualized in Verify after being parsed and processed. This view facilitated faster, easier and more accurate clinical data validation. Upcoming developments will also allow users to validate, query and edit the contents of tables directly.

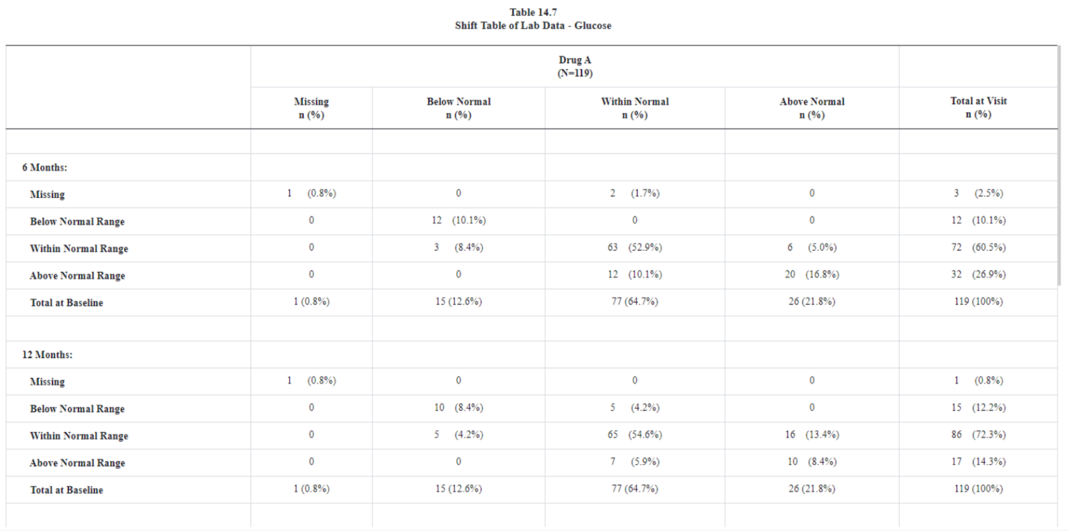


Table 14.7
Shift Table of Lab Data - Glucose

	Drug A (N=119)				
	Missing n (%)	Below Normal n (%)	Within Normal n (%)	Above Normal n (%)	Total at Visit n (%)
6 Months:					
Missing	1 (0.8%)	0	2 (1.7%)	0	3 (2.5%)
Below Normal Range	0	12 (10.1%)	0	0	12 (10.1%)
Within Normal Range	0	3 (8.4%)	63 (52.9%)	6 (5.0%)	72 (60.5%)
Above Normal Range	0	0	12 (10.1%)	20 (16.8%)	32 (26.9%)
Total at Baseline	1 (0.8%)	15 (12.6%)	77 (64.7%)	26 (21.8%)	119 (100%)
12 Months:					
Missing	1 (0.8%)	0	0	0	1 (0.8%)
Below Normal Range	0	10 (8.4%)	5 (4.2%)	0	15 (12.2%)
Within Normal Range	0	5 (4.2%)	65 (54.6%)	16 (13.4%)	86 (72.3%)
Above Normal Range	0	0	7 (5.9%)	10 (8.4%)	17 (14.3%)
Total at Baseline	1 (0.8%)	15 (12.6%)	77 (64.7%)	26 (21.8%)	119 (100%)

Table 4: Table analysis view on Verify's web platform.

CONCLUSION

In conclusion, Beaconcure has developed a unique data processing pipeline which allows storing table values along with header values to provide contextual information about cells. This data mining process opens the door for more efficient data validation, cross referencing and cross-table checks, dynamic editing and version control.

CONTACT INFORMATION

Name: Ilan Carmeli
Company: Beaconcure Ltd
Work Phone: +972-54-3183009
E-mail: ilan@beaconcure.com
Website: www.beaconcure.com