Paper DV06

An Approach to Dynamic Data Visualization with dplyr, DT and Shiny

Nicholas Masel, Janssen Research & Development, Raritan NJ, United States
Megan Kendellen PhD, Janssen Research & Development, Titusville NJ, United States
Richard Ryan PhD, Janssen Research & Development, Titusville NJ, United States

ABSTRACT

Timelines to develop regulatory documents are ever shortening, making manual review of large statistical outputs increasingly untenable. Manual review is also error-prone and hinders collaboration with cross-functional subject matter experts to select relevant data for in-text presentation. To address these challenges, we have developed an automated Data Visualization Tool. The tool converts data table outputs in Rich Text Format (.RTF) or .DOCX to a data frame and determines the number of columns within the table. Coupled with dplyr's _at() variants, this provides a solution to dynamically process any table regardless of the number of columns. A Shiny user interface makes the tool end-user friendly. The result is a flexible tool used to rapidly identify data points that exceed any desired incidence or are differentially reported across treatment groups at any desired threshold. The tool increases accuracy, reduces quality control burden, and allows for easy manipulation of thresholds to facilitate identification of trends during document development.

INTRODUCTION

Data from clinical trials are complex and voluminous and will become increasingly so as regulatory requirements increase. Clinically-sound interpretation of these data, and their clear and concise presentation in regulatory documents, are critical to the success of clinical development programs. Due to their role in authoring regulatory documents, regulatory medical writers (RMWs) are typically responsible for preliminary efforts to review clinical data as part of preparing the draft text. RMWs may also drive discussions with cross-functional subject matter experts to delineate the appropriate clinical interpretation of study data.

Efficacy and pharmacokinetic data lend themselves to pre-specified statistical analysis with regimented potential outcomes. However, the interpretation and presentation of safety data (eg, adverse events [AEs]) are complicated by their more subjective nature, and the fact that safety outputs may extend over dozens of pages. Historically, RMWs manually reviewed these lengthy safety outputs to identify and summarize potentially clinically relevant AEs in text. This laborious process is time-consuming, prone to error, and limits the ability of the RMW to raise potentially important trends to clinicians and other subject matter experts in a timely fashion.

To meet the current demands for shorter timelines and higher quality from Health Authorities, as well as to empower RMWs to support clinical discussions, we developed a web-based Data Visualization Tool. Currently, the intended users are RMWs, with additional beneficiaries including clinicians and programming personnel.

DATA VISUALIZATION TOOL

In a collaboration between RMWs and the Analytics Innovation Team, we created a web-based Data Visualization Tool to facilitate review and interpretation of safety table outputs. Specifically, the tool automates the identification of AEs that exceed any desired incidence or that are differentially reported across treatment groups at any threshold. While the tool is primarily intended to review safety data, it can also be used with outputs with similar structure/formatting (eg, medical history and concomitant medications). The tool could be used to support preparation of any document in which safety data are a key component, including Clinical Study Reports, Investigator Brochures, Risk Management Plans, and Development Safety Update Reports.

The expected benefits of the tool are as follows:

- **Rapid and efficient identification of frequently reported and differentially reported AEs**. The tool provides powerful flexibility and can be tailored to the individual needs of a particular data set due to the easily-manipulated thresholds.
    - The tool can accommodate a variety of output formats and any version of the Medical Dictionary for Regulatory Activities (MedDRA).
    - Data may be reviewed at either the system organ class (SOC) or preferred term level (or both together), depending on the needs of the user.
    - Data can be reviewed at the incidence (ie, frequency) level or at the subject level. The latter may be especially pertinent for rarer AEs that have special clinical import (ie, SAEs or AEs of clinical interest).
    - The tool includes a user-friendly interface discussed in detail below.
- **Elimination of manual output review, which is one of the most time-intensive and error-prone steps in the development of regulatory documents**. We expect that this tool will facilitate an overall reduction in the time need to prepare data-driven regulatory documents for Health Authorities.
- **Provision of an overall visualization of data using rudimentary heat mapping**. For each AE, the tool color codes the treatment group(s) that have a higher or lower incidence than the selected comparator treatment group. Any treatment group may be identified by the user as the comparator. The tool can accommodate outputs with any number of treatment groups.
- **Early engagement with clinicians and pertinent subject matter experts** due to the automated and flexible manner in which data may be queried. The potential for earlier identification of data trends may reduce the likelihood of time-consuming ad hoc outputs being produced late in the document development process.
    - It should be noted that, as a non-GxP tool, it is intended to enhance the role of the RMW in the identifications of data trends and will not be used to produce outputs for inclusion in regulatory documents.
- **Improved data accuracy** of working drafts of documents to be reviewed by functional teams.
- **Reduced quality control burden** by promoting higher fidelity in documents as written by the writer, and providing a tailored source to support QC reviews, if desired.

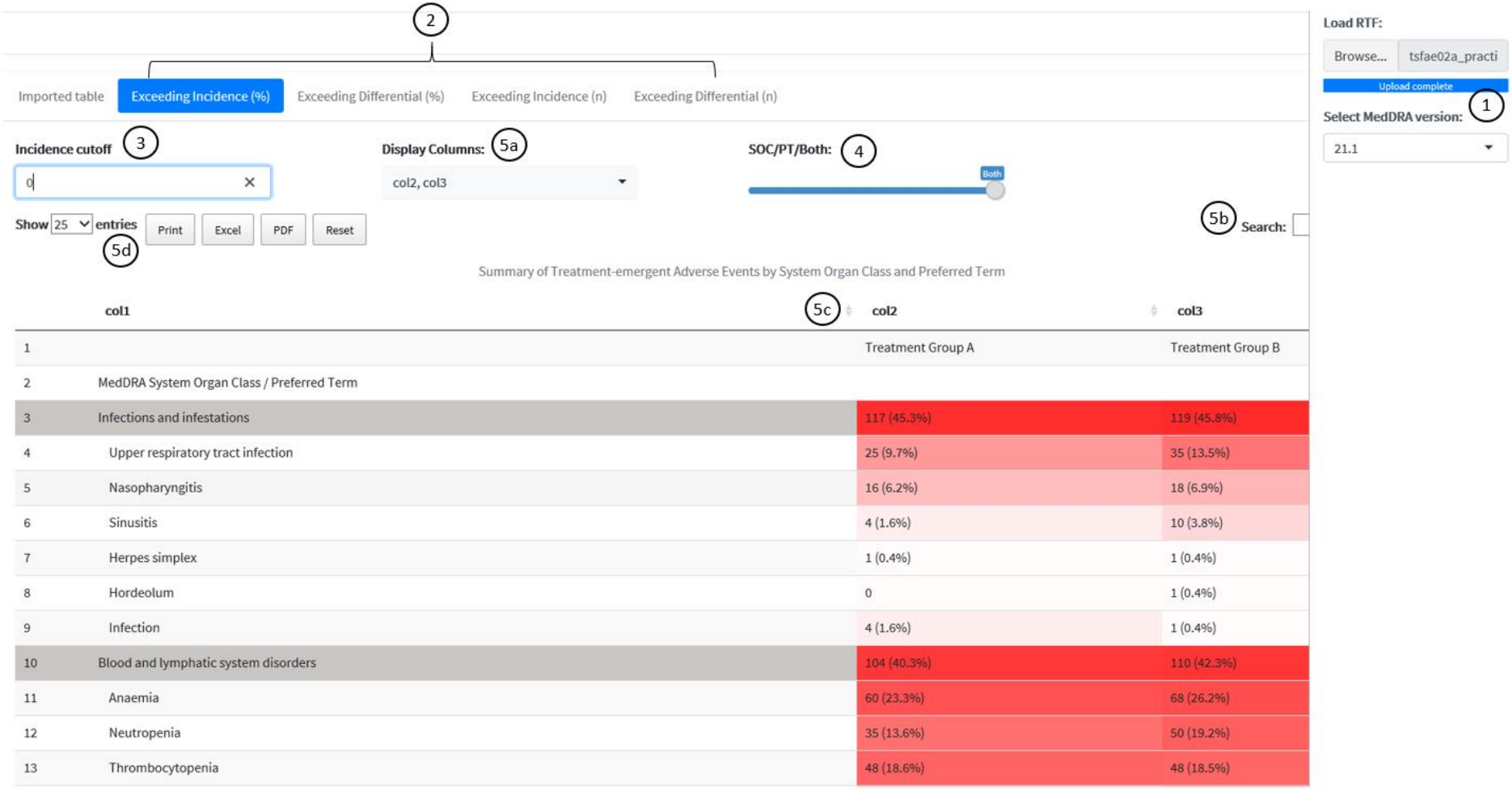
It is expected that Data Visualization Tool will enhance the value RMWs can contribute to the development of regulatory documents in terms of quality and content and accelerate the preparation of these documents. The Data Visualization Tool is now in use across the RMW global team at Janssen Research & Development and will continued to be improved in usability and functionality.

USER INTERFACE AND USE OF THE DATA VISUALIZATION TOOL

The Data Visualization Tool has a simple and approachable user interface. The following steps are followed to review data (Figure 1):

1) Load the desired .RTF or .DOCX file and select the relevant version of MedDRA.
2) Select appropriate tab for desired analysis among the following options:
    a. Data may be queried by incidence or by number of subjects
    b. Data may be queried by group (ie, assess frequency within each treatment group to highlight those exceeding a desired threshold) or among groups (ie, compare frequency across treatment groups to highlight those that differ among groups at a desired threshold)
        i. The tool defaults to comparing all treatment groups to the left-most column in the output (termed the comparator for the purposes of this paper). See more information below about how to manipulate which treatment group is the comparator.
3) Enter the desired threshold in the appropriate free-response field (see options below). Note that results for each query (ie, incidence and difference in incidence) are produced by the tool independently.
    a. Desired cutoff of incidence (in any treatment group)
    b. Desired cutoff for difference in incidence (among treatment groups)
    c. Desired cutoff of number of subjects (in any treatment group)
    d. Desired cutoff for difference in number of subjects (among treatment groups)
4) If applicable, select whether the tool should display AEs by SOC, preferred term, or both.
    a. SOC and preferred term(PT) also automatically differentiated by grey coloration and indentation.
5) As desired, the user may:
    a. Select a subset of treatment groups to be displayed
    b. Search for individual terms included in results
    c. Sort any treatment group column by increasing or decreasing incidence, quickly ordering the terms of interest so that the user may present them accordingly in a regulatory document.
    d. Export results in Excel or PDF formats.

Figure 1: User Interface of the Data Visualization Tool



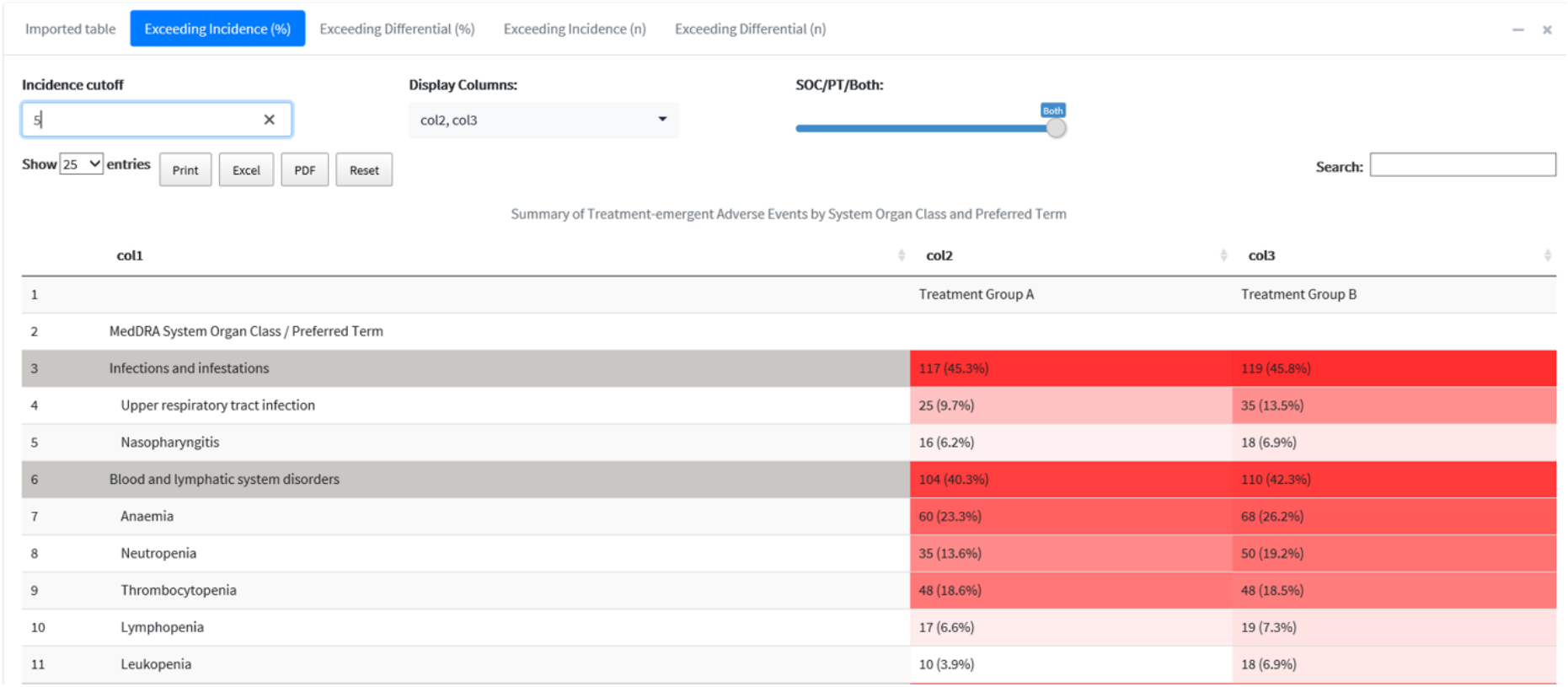
The user may repeat these steps for as many iterations as are desired across the 4 output tabs and for various thresholds. Each iteration typically takes no more than 30 seconds (less, if the user does not have to re-load the source file) and replaces 10 to 15 minutes of work in the manual process.

The above steps automate the process by which a user can identify, for example, AEs that exceed a clinically relevant threshold in any treatment group (eg, most frequently reported AEs; Figure 2) or that are differentially reported among treatment groups (eg, AEs that may be associated with study drug more or less frequently than a placebo; Figure 3, in which results are shown only for preferred terms). This information is typically an essential part of regulatory documents, as the sponsor highlights the most important safety information for evaluation by Health Authorities.

Importantly, the tool also provides an overall visualization of the data in the following ways:

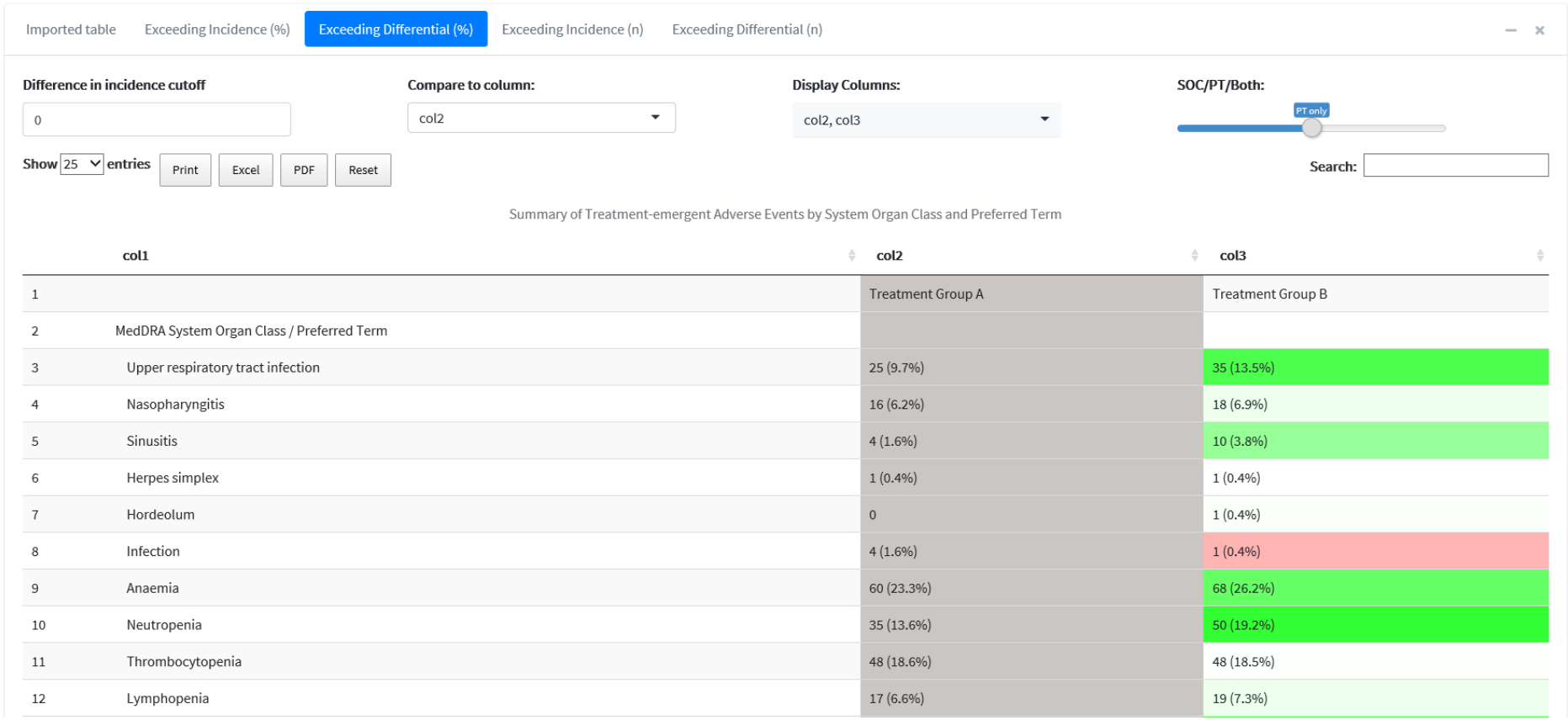
1) The tool indicates the treatment group(s) in which the incidence or number of subjects meets or exceeds the desired threshold by color. The intensity of the color is correlated to the degree by which that data exceeds the threshold compared with other such terms (ie, heat mapping). This information adds value, for example, by quickly showing the user whether terms tend to cluster near the chosen threshold (and thus a more informative threshold can be queried), whether a few AEs are reported at an unexpectedly high frequency compared to other observed AEs, or whether the most frequent AEs cluster in a particular SOC.
2) In the tabs showing a difference in the incidence or number of subjects among treatment groups, two-tone heat mapping is utilized to indicate whether the value for a particular treatment group is greater (ie, green) or less (ie, red) than the comparator group. This information adds value, for example, by quickly showing the user whether AEs are more frequently reported in a placebo group or a treatment group.

Figure 2: Example Results for Most Frequently Reported AEs



	col1	col2	col3
1		Treatment Group A	Treatment Group B
2	MedDRA System Organ Class / Preferred Term		
3	Infections and infestations	117 (45.3%)	119 (45.8%)
4	Upper respiratory tract infection	25 (9.7%)	35 (13.5%)
5	Nasopharyngitis	16 (6.2%)	18 (6.9%)
6	Blood and lymphatic system disorders	104 (40.3%)	110 (42.3%)
7	Anaemia	60 (23.3%)	68 (26.2%)
8	Neutropenia	35 (13.6%)	50 (19.2%)
9	Thrombocytopenia	48 (18.6%)	48 (18.5%)
10	Lymphopenia	17 (6.6%)	19 (7.3%)
11	Leukopenia	10 (3.9%)	18 (6.9%)

Figure 3: Example Results for Differentially Reported AEs by Preferred Term



Imported table | Exceeding Incidence (%) | Exceeding Differential (%) | Exceeding Incidence (n) | Exceeding Differential (n)

Difference in incidence cutoff: 0

Compare to column: col2

Display Columns: col2, col3

SOC/PT/Both: PT only

Show 25 entries | Print | Excel | PDF | Reset

Search:

Summary of Treatment-emergent Adverse Events by System Organ Class and Preferred Term

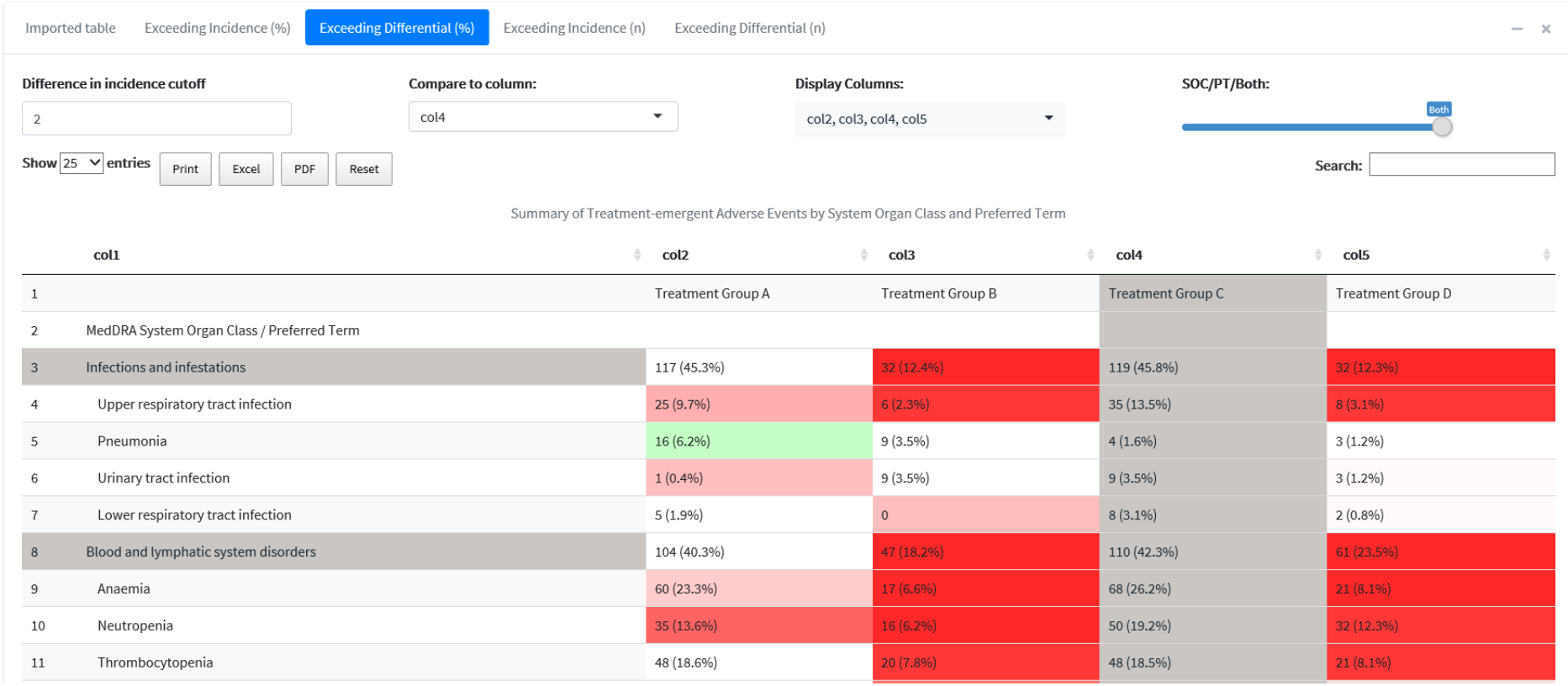
	col1	col2	col3
1		Treatment Group A	Treatment Group B
2	MedDRA System Organ Class / Preferred Term		
3	Upper respiratory tract infection	25 (9.7%)	35 (13.5%)
4	Nasopharyngitis	16 (6.2%)	18 (6.9%)
5	Sinusitis	4 (1.6%)	10 (3.8%)
6	Herpes simplex	1 (0.4%)	1 (0.4%)
7	Hordeolum	0	1 (0.4%)
8	Infection	4 (1.6%)	1 (0.4%)
9	Anaemia	60 (23.3%)	68 (26.2%)
10	Neutropenia	35 (13.6%)	50 (19.2%)
11	Thrombocytopenia	48 (18.6%)	48 (18.5%)
12	Lymphopenia	17 (6.6%)	19 (7.3%)

REVIEW OF DIFFERENTIAL INCIDENCE IN STUDIES WITH 3 OR MORE TREATMENT GROUPS

For statistical outputs that include 3 or more treatment groups, the user may select any of the desired treatment groups as the comparator for all other treatment groups. This information adds value by allowing the user to quickly query whether various treatment groups differ from each other in any desired combination (Figure 4, in which Treatment Group C is used as the comparator, and Figure 5, in which Treatment Group D is used as the comparator and Treatment Group A has been removed from comparison). After selection of the desired comparator, the remaining steps are performed as described above.

As described above, a user can utilize this Data Visualization Tool to answer many of the most common queries for safety data sets when preparing regulatory documents (like the examples noted above). Most importantly, these queries are performed with unmatched speed, flexibility, and accuracy that is impossible to achieve with the historical manual review process.

Figure 4: Example Results for Differentially Reported AEs Among 3+ Treatment Groups



Imported table | Exceeding Incidence (%) | Exceeding Differential (%) | Exceeding Incidence (n) | Exceeding Differential (n)

Difference in incidence cutoff: 2

Compare to column: col4

Display Columns: col2, col3, col4, col5

SOC/PT/Both: Both

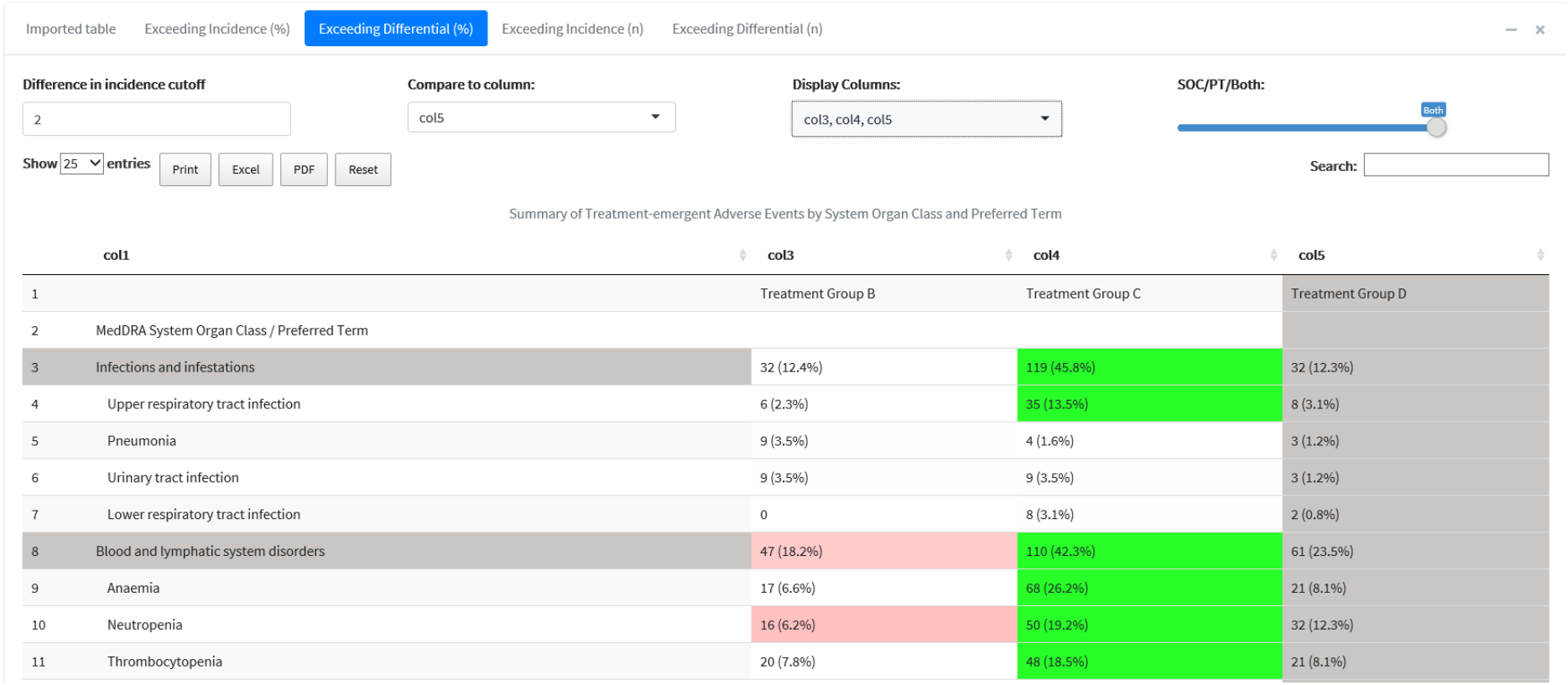
Show 25 entries | Print | Excel | PDF | Reset

Search:

Summary of Treatment-emergent Adverse Events by System Organ Class and Preferred Term

	col1	col2	col3	col4	col5
1		Treatment Group A	Treatment Group B	Treatment Group C	Treatment Group D
2	MedDRA System Organ Class / Preferred Term				
3	Infections and infestations	117 (45.3%)	32 (12.4%)	119 (45.8%)	32 (12.3%)
4	Upper respiratory tract infection	25 (9.7%)	6 (2.3%)	35 (13.5%)	8 (3.1%)
5	Pneumonia	16 (6.2%)	9 (3.5%)	4 (1.6%)	3 (1.2%)
6	Urinary tract infection	1 (0.4%)	9 (3.5%)	9 (3.5%)	3 (1.2%)
7	Lower respiratory tract infection	5 (1.9%)	0	8 (3.1%)	2 (0.8%)
8	Blood and lymphatic system disorders	104 (40.3%)	47 (18.2%)	110 (42.3%)	61 (23.5%)
9	Anaemia	60 (23.3%)	17 (6.6%)	68 (26.2%)	21 (8.1%)
10	Neutropenia	35 (13.6%)	16 (6.2%)	50 (19.2%)	32 (12.3%)
11	Thrombocytopenia	48 (18.6%)	20 (7.8%)	48 (18.5%)	21 (8.1%)

Figure 5: Example Results for Differentially Reported AEs Among a Subset of 3+ Treatment Groups



Imported table | Exceeding Incidence (%) | Exceeding Differential (%) | Exceeding Incidence (n) | Exceeding Differential (n)

Difference in incidence cutoff: 2

Compare to column: col5

Display Columns: col3, col4, col5

SOC/PT/Both: Both

Show 25 entries | Print | Excel | PDF | Reset

Search:

Summary of Treatment-emergent Adverse Events by System Organ Class and Preferred Term

	col1	col3	col4	col5
1		Treatment Group B	Treatment Group C	Treatment Group D
2	MedDRA System Organ Class / Preferred Term			
3	Infections and infestations	32 (12.4%)	119 (45.8%)	32 (12.3%)
4	Upper respiratory tract infection	6 (2.3%)	35 (13.5%)	8 (3.1%)
5	Pneumonia	9 (3.5%)	4 (1.6%)	3 (1.2%)
6	Urinary tract infection	9 (3.5%)	9 (3.5%)	3 (1.2%)
7	Lower respiratory tract infection	0	8 (3.1%)	2 (0.8%)
8	Blood and lymphatic system disorders	47 (18.2%)	110 (42.3%)	61 (23.5%)
9	Anaemia	17 (6.6%)	68 (26.2%)	21 (8.1%)
10	Neutropenia	16 (6.2%)	50 (19.2%)	32 (12.3%)
11	Thrombocytopenia	20 (7.8%)	48 (18.5%)	21 (8.1%)

A GUIDE TO WHAT'S HAPPENING SERVER-SIDE

The application is written in R, utilizing the Shiny package to create the application. For those unfamiliar with Shiny, the basic concept is a user interface talks to a server that processes and reacts to a user's actions and updates the interface accordingly. For more information regarding Shiny, please refer to the recommended reading section below.

PROCESS RTF AND/OR DOCX

As previously outlined, the workflow begins with loading .RTF or .DOCX files to the web-based Data Visualization - application. The application then reads the RTF as a character vector where each row is 1 element. Within the element, columns are separated with a pipe, which allows us to split 1 variable into multiple variables, and we are left with a data frame of rows and columns to manipulate however desired.

Each column is processed to determine which of these contain count and percentages. The application is expecting an "XX (XX.X%)" structure, or something similar, and will count rows that contain either open/closed parentheses or a percent sign. If a certain threshold of rows is met that contain these characters, it will be included for further processing.

COUNT AND PERCENTAGE VARIABLE GROUPS

After the application has identified which columns contain counts and percentages, they are further split out into a set of variables that contain counts and a set of variables that contain percentages.

Using the dplyr _at variances, a character vector of column names to perform tasks across multiple columns is employed. Notably, this activity is possible even if the column names are unknown, allowing data processing to remain dynamic regardless of the number of columns (Figure 6).

Figure 6: dplyr _at Variances for Dynamic Variable Creation

```
# Create count variables
counts <- df_with_columns %>%
  select_at(columns_with_percentages$col) %>%
  mutate_at(columns_with_percentages$col, find_count) %>%
  rename_at(columns_with_percentages$col, rename_count)
```

Here, a new set of variables that contain the count is created. We can use an object that contains the columns that were identified as having counts and percentages in previous steps to select, mutate to the desired result, and name as appropriate. This new set of count variables can now be used in a similar fashion for further processing downstream.

After the count and percent variable groups are obtained, user can set a threshold from the user interface to restrict records shown to only those where at least 1 column in the table meets the threshold requirements.

HEAT MAP

As with the previous summary of count and percentages, these are all stored as groups of variables so these groups can be used accordingly with the _at dplyr function variances. Coupled with functionality from the DT package, the color and break points are set to format the table to these defined color intervals.

The same logic could be applied as Figure 6, using a select_at, mutate_at, and rename_at, or mutate_at to also set the variable names, eliminating the need for rename_at (Figure 7).

Figure 7: dplyr mutate_at for Dynamic Variable Creation

```
# subject count cutoff variables
count_cutoffs <- counts %>%
  mutate_at(vars(starts_with("cnt_")), 1
            .funs = list(
              eeis_heatmap = ~ . - as.integer(input$sli_subject_cutoff)
            )
           )
```

As shown in step 1 of Figure 7, mutation across all variables that start with “cnt_” using the starts_with helper function is performed. This eliminates the step of defining a new object to store the list of count variables. In step 2, the function within the mutate_at is defined since this is a one-time use function. This function will create a new variable with “_eeis_heatmap” appended to the end of each variable in step 1. A variable named cnt_col1 will now have a corresponding variable named “cnt_col1_eeis_heatmap” that will contain the difference in “cnt_col1” and the user specified cutoff or threshold selected within the user interface of the application. Once again, the process is independent of how many variables are present. The entire set of variables that start with “cnt_” is processed, and a new set of variables suffixed with “_eeis_heatmap” is created. Moving forward, this new set of variables using the ends_with(“_eeis_heatmap”) can be referenced with a helper function. This process can be applied to all derivations needed, thus creating an application that is dynamic, adjusting relative to the file loaded and the number of columns that contain counts and percentages within the file.

CONCLUSION

The Data Visualization Tool solves a number of challenges inherent in the method by which RMWs have historically reviewed safety outputs and prepared regulatory documents. In doing so, the tool offers both faster timelines and increased quality that we hope will be a considerable benefit to the clinical development process.

ACKNOWLEDGMENTS

We would like to acknowledge the contributions of the following individual and thank them for invaluable input: Satish Murthy, Sumesh Kalappurakal, Neil Garrett, Frank Meloni, Lin Li, Mary Ost, Ruth Smans, Eveline Hoste, Sharon Ramella, and Jonathan Zubkoff.

RECOMMENDED READING

https://shiny.rstudio.com/
https://dplyr.tidyverse.org/reference/mutate_all.html
https://rstudio.github.io/DT/010-style.html

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:
Janssen Research & Development
1125 Trenton-Harbourton Road
Titusville NJ 08560
Email: nmasel@its.jnj.com; mkendel1@its.jnj.com; rryan3@its.jnj.com;