PharmaSUG 2014 - Paper DS12

Effective Use of Metadata in Analysis Reporting

Jeffrey Abolafia, Rho Inc., Chapel Hill, NC USA

ABSTRACT

Many organizations are effectively using metadata for the creation and validation of clinical and analysis databases. The use of metadata for analysis reporting, however, lags behind that of database production. While many sponsors have submitted define.xml to document ADaM databases, very few of these files have included analysis results metadata. This despite a recent CDISC pilot, which demonstrated that results level metadata adds significant value to a regulatory submission.

As the CDISC ADaM model has matured as an analysis dataset standard it, combined with results related metadata, can be utilized to facilitate producing displays and statistical analysis and easily be extended to generate the results portion of define.xml file. This presentation will examine how the ADaM standard used in conjunction with results level metadata can be used to generate statistical reports more efficiently, substantially add value and traceability to the define file, and facilitate the management and tracking of analysis tasks.

INTRODUCTION

It's worthwhile reviewing why metadata systems have become so popular and what problems these systems solve. Many organizations are effectively using metadata to create and validate clinical and analysis datasets. Organizations typically develop metadata at the dataset and variable levels in order to create datasets more efficiently. A review of the most recent PhUSE, PharmaSUG, and SAS® Global Forum proceedings turns up a plethora of presentations on metadata systems created to manage datasets and variable level metadata (including several by the author). Given the current submission requirements, one could argue that datasets and variable level metadata has become a *de facto* requirement for clinical studies.

In recent years the FDA has clearly stated its preference for receiving both clinical and analysis data that conform to CDISC standards. As a result, CDISC models have become the *de facto* standard for submitting data to the FDA. This, in turn, means that sponsors have begun to produce CDISC-compliant databases in order to meet the FDA's submission requirements.

In the short term this has led to additional work and higher costs. Introducing CDISC models, especially SDTM, has a significant effect on work streams, work flow, and work processes[(1)]. Adding SDTM to the work flow means adding an entirely new work stream. The flow of work is no longer from the Data Management System (DMS) to analysis datasets; it is now from the DMS *to SDTM* to analysis datasets. Furthermore, the metadata requirements for creation and documentation of SDTM and ADaM deliverables are significantly greater than their pre-CDISC counterparts. Metadata must be specified at the dataset, variable, and value level. This affects timelines, budget, and resources. In general, more of everything is needed.

THE BUSINESS CASE FOR METADATA

An earlier paper by the author (see "References," below) demonstrated that well constructed metadata and metadata access tools can significantly improve the creation of the datasets and documents that comprise an electronic submission. It outlined how metadata-driven applications and utilities can be integrated into standard business processes, speeding the production and improving the quality of deliverables.

The paper looked at how dataset and variable-level metadata is utilized throughout a clinical study. First, regardless of whether one is creating datasets conforming to CDISC or proprietary standards, content specification is typically stored as machine readable metadata. This is vital to ensure that the standard is maintained and made available throughout the life cycle of a project. Second, the metadata is extended to include programming specifications and other documentation for creating the datasets. Third, the same metadata is used as input for the programs that create clinical and analysis datasets. Fourth, the datasets produced are checked against the metadata to make sure that they are standard-complaint. Fifth, the metadata is enhanced so that it includes all of the information needed to produce documentation (aka the "define file") for the submission's datasets. Finally, the metadata can be used to produce specifications or reports in a variety of formats for a variety of audiences. In summary, dataset and variable metadata is used as a single source for efficiently documenting and creating datasets, as well as providing the traceability of the data from collection to submission required by regulatory agencies.

EXTENDING METADATA TO DISPLAYS AND ANALYSES

The end-to-end benefits of metadata-based systems are clear. What is not as obvious is why the use of metadata for analysis reporting lags behind that of database production. Recent PhUSE and PharmaSUG conferences have been inundated with papers describing the use of metadata for producing CDISC-compliant data and documentation. There have, however, been very few papers on metadata systems for producing and documenting statistical results. Only a handful of sponsors have submitted a define file that includes results-level metadata. This is surprising, since: statistical displays and analyses are a key deliverable for clinical studies; the functionality of metadata for datasets can be easily extended to analysis reporting; and because traceability should start with *results* not *analysis datasets*.

Standard output for most studies includes a series of tabulations, graphic displays, and listings of individual observations, collectively referred to as TFLs. While a study can require hundreds of TFLs, it is the norm for these displays to be based on a much smaller number of unique layouts. Twenty tables could have similar layouts, for example, varying only by the population used in each table – treated patients, age greater than 65, and so on. With well designed metadata, a single program shell can produce all 20 tables, and can even be generalized to produce similar tables for other studies.

Many displays are not only standard across studies but also across therapeutic areas. Specifications for these displays can be stored as metadata in global metadata libraries. The information in these displays that varies across studies can be stored as display level metadata in project specific libraries. This metadata can be used as input to the programs that create displays. More importantly, results-level metadata can add significant value to a submission. This was demonstrated in the updated CDISC Pilot Project [(2)]. Display level metadata can easily be broadened to include all of the data needed for the results portion of the define file for analysis datasets. As with dataset/variable metadata, results level metadata can serve as a single source for documentation and producing displays and analyses and provide the input needed for the define file. In addition, results level metadata provides the traceability required by the FDA: it documents results and provides the regulatory reviewer with the means to trace results back to their source (programs, datasets, and statistical analysis plans).

USING RESULTS-LEVEL METADATA

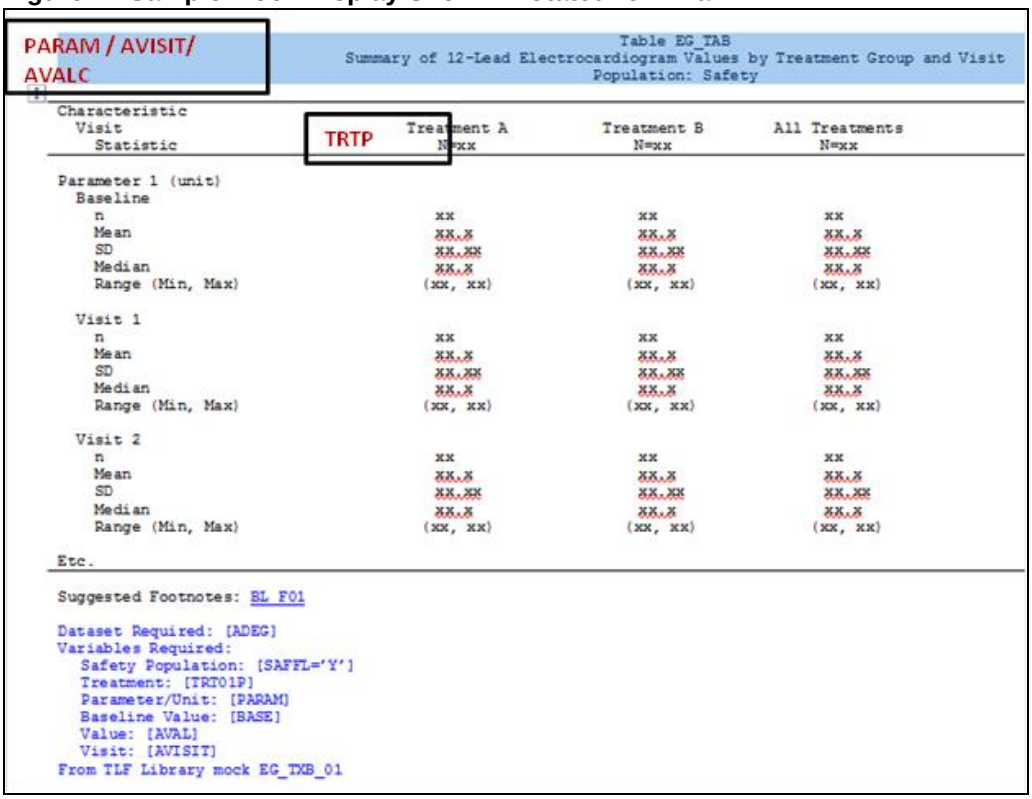
In this section we describe the use of results level metadata for: producing displays, creating the results component of the define file, and managing/tracking displays and analysis. In the process we examine the metadata architecture and related components required to produce these deliverables.

TFL LIBRARY

To get started we need a TFL library consisting of standard displays that can be used for most studies. Due to their relative uniformity, safety display shells predominate in the TFL library. For efficiency, these

displays should be annotated to a standard like ADaM (See **Figure 1**). The second component required is a library of generic validated programs which will produce the displays. The third piece is the display (or results) level metadata which will contain data about each display or analysis. The final item needed is a set of tools or macros to seamlessly read the display level metadata and provide it to the display creating programs.

Figure 1. Sample Mock Display Shell Annotated for ADaM



DISPLAY METADATA

Most of the information describing a display will be contained in the "DISPLAY" table. The structure of this table is one record per TFL. Each record will contain specific documentation about a single display (or analysis). **Figure 2,** next page, provides an example of display level metadata.

Figure 2. Sample Display Level Metadata

Number	TableLetter	Title	Population	Population_Label	Selection_Criteria	Footnotes	Dataset
8.7.3.1	DM_TAA	Demographics Characteristics	ITTFL = 'Y'	Intent to Treat Population		ITT5, FT30, FT31	ADSL
8.7.3.2	AE_TAB	Adverse Events by Treatment Group	SAFFL = 'Y'	Safety Population		PPP5, FT30	ADAE
8.7.4.1	VS_TAB	Vital Sign Values by Treatment Group	SAFFL = 'Y'	Safety Population		FT1,FT2, FT3, SRC5	ADVS

Outcome	Treatment	TreatmentValues	TimeVariab	TimePoints	BaselineV	Analysis Variables
	TRT01P	1,2				AGE,WEIGHTB, HEIGHTB
	TRT01P	1,2				ANYAE, AEBODSYS, AEDECOD
AVAL	TRT01P	1,2	AVISIT	1,2,3,4,5	BASE	AVAL

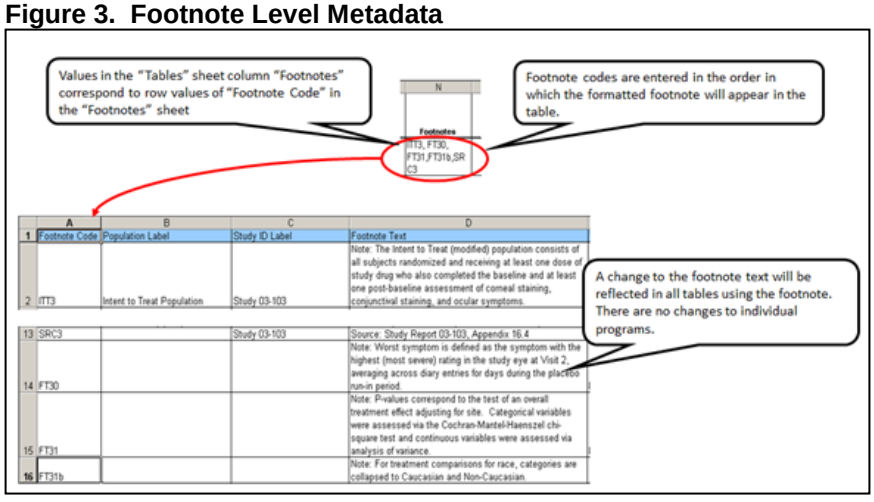
The DISPLAY metadata describes key features of each TFL including:

- Display number
- Title lines
- Footnote codes
- Datasets used by the table
- Display type (Table, Figure, Listing)
- Population criteria
- Filtering information expressed as both descriptive text to use in titles and syntactically valid SAS Program statements
- A list of variables needed to create the display
- Treatment and time point information

FOOTNOTE METADATA

The FOOTNOTES table (**Figure** 3) complements DISPLAY metadata. It contains one record per unique footnote and consists of a field with a short footnote code and a longer text field containing the actual footnote text .The linkage of the DISPLAY and FOOTNOTES tables emphatically demonstrates the

Figure 3. Footnote Level Metadata



power of metadata-driven processes. A given footnote is likely to be used in multiple displays. Without metadata a footnote's text has to be manually changed in multiple programs. Using metadata, the task is vastly easier and the output more reliable: a single change is made to text in the FOOTNOTE table. The affected programs can be rerun without modification, since all the changes are contained in the metadata and passed to the programs using validated tools.

PROJECT-LEVEL METADATA

The third and final metadata table required for producing TFLs (and the results portion of the define file) is project-level metadata. This table is structured as one record per project and contains information specific to a project or drug development program. The project table includes, but is not limited to:

- Name and description of study
- Location on network of study
- Location of study database, format, and macro libraries
- Other directory locations
- SAS Options
- Treatment group information
- Visit information
- Sponsor or project specific preferences for displays
    - Styles
    - Formats

SEAMLESS ACCESS TO THE METADATA

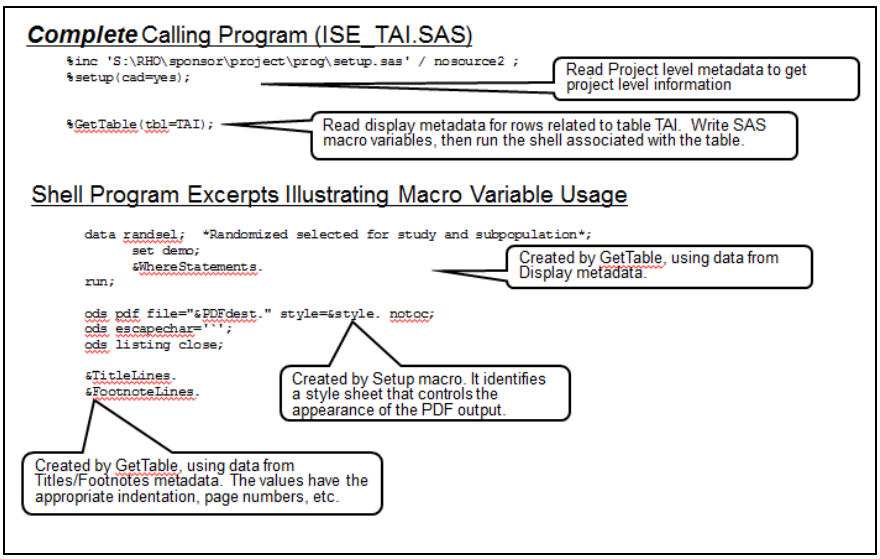
While developing our metadata system one of the early lessons learned was the need to make access to the metadata more or less transparent. To underscore the need for access tools, consider the programming requirements needed for accessing metadata that describes a display such as a statistical table. The table number must be located in the DISPLAY metadata. We must also gather and correctly sequence the footnotes from the FOOTNOTES table. Finally, we must retrieve standard headers and footers from the GLOBAL table. Once all the pieces are identified, they need to be presented to the table-writing program in an agreed-upon format (macro variables, datasets, etc.).

To be thorough, we should add checks to ensure data quality – are all the footnotes in the DISPLAY table actually present in the FOOTNOTES table? Do we have complete title text? And so on.

We could, of course, write code in each table program to perform these actions. More likely, we would want a tool that would do the work for us, reading the required tables, and creating a set of macro variables that would make the metadata readily accessible. Two of the tools we created that are especially relevant for display production are: 1) the “setup” macro, which retrieves information from the project level metadata and makes it available to the display program; and 2) the “getspecs” macro, which retrieves display specific data from the display level metadata and passes it to the display program. The bottom portion of **Figure 4** (next page) demonstrates how these two tools are used by a generic display program.

Results level metadata supplemented with and a set of tools and generic display programs can significantly increase the efficiency and quality of display production. The top portion of Figure 4 provides an example of the ease of creating a display with a metadata based system.

Figure 4. Sample Display Program



DEFINE FILE PRODUCTION

The recent CDISC pilot noted that results level metadata adds significant value to a regulatory submission and describes, via an ODM schema extension, the contents of the metadata. The metadata described above provides almost all of the input required to create the results part of the define file. Only a few additional fields have to be added to the display metadata table. Figure 5 provides an example of the results section of the define file for a single display.

Figure 5. Results Component of Define File for ADaM

Display	Table 14-3.12 Mean NPI-X total score from Weeks 4 through Week 24 - Windowed (Efficacy Population)
AnalysisResult	**Descriptive statistics (n, mean, standard deviation, median, min, and max)**
Analysis Parameter(s)	NPTOTMN=Mean NPI-X (9) Total (Week 4 to 24)
Analysis Variable(s)	AVAL BASE
Reason	pre-specified in SAP
Data References (incl. Selection Criteria)	ADQSNPIX [EFFFL="Y" and PARAMCD="NPTOTMN" and AVISIT="Weeks 4-24" and ANL01FL="Y"]
Documentation	SAP Section 10.2
Programming Statements	PROC UNIVARIATE data= ADQSNPIX (where =(EFFFL = 'Y' and PARAMCD = 'NPTOTMN' and AVISIT='Weeks 4-24' and ANL01FL= 'Y')); Var BASE AVAL; By TRTPN; Run;

"Updated SDTM/ADaM pilot Package", online at www.cdisc.org/sdtm-adam-pilot-project

Figure 6, next page, lists the fields that are needed to produce the results module of the define file. We can see that six of the ten fields are already used to generate displays, one field may have been used to create displays, and there are three new fields to populate that are not used in TFL production.

Figure 6. Fields in the Results Component of the Define File

Variable	Display Production
Display Identifier	Yes
Display Name	Yes
Analysis Result	No
Analysis Parameters	Yes
Analysis Variables	Yes
Reason	No
Dataset	Yes
Selection Criteria	Yes
Documentation	No
Programming Statements	Maybe

Once the metadata described above is populated the program or script used to produce the define file can be extended to generate the results component. At this point in time, an out of the box tool to produce the results section of the define file with the schema-compliant level of detail is not available.

MANAGEMENT/TRACKING

One of the many benefits of a metadata based system is that it is inherently multi-use. A metadata source table can easily be used for multiple types of tasks throughout the project life cycle. Initially our display level metadata was designed to produce a large number of displays for submission projects more efficiently. Most of our submission projects included both an integrated summary of safety and an integrated summary of efficacy, often comprised of more than 500 displays. As a result, we needed a tool that could manage the production, validation, and status of a large number of displays. We soon figured out that the same system used to generate displays could also serve as a management/tracking tool for display production.

The TASK metadata table was added to facilitate managing display production. The structure of this table is one record per display per task per study. The TASK table can be linked to the DISPLAY table by table number or table ID. Fields in the TASK table include:

- Table ID
- Table Number
- Task (create program, create validation program, validate cosmetics, validate statistics, QC, revise program, and verify revision)
- Status (not started, started, completed)
- Owner
- Date started
- Date completed
- Due date
- Comments (or instructions)

Figure 7, next page, shows an excerpt of the TASK table. It demonstrates that for a given display one can determine if: a display is assigned for programming; programming is complete and awaits validation; validation has started or has been completed; issues were found during the validation process; the statistics and cosmetics have been validated; and QC has been performed. Since all of this information is

stored in a database, a variety of reports can be created to show the overall status of display production. The system also provides documentation that all displays have been validated and undergone QC.

Figure 7. TASK Metadata table

Name	Number	Task	Status	Owner	Created	Started	Completed	Next Due Date	Instructions
TAA_DS_01	14.1.1	Create Program	Completed	kreece	08-Oct-2012 09:37:36 AM	11-Oct-2012 11:12:46 AM	11-Oct-2012 11:12:47 AM		
TAA_DS_01	14.1.1	Create Validation Program	Completed	pnguyen	08-Oct-2012 09:37:36 AM	16-Oct-2012 04:24:23 PM	16-Oct-2012 04:24:23 PM		
TAA_DS_01	14.1.1	Validate Cosmetics	Completed (Pass)	rwoolson	08-Oct-2012 09:37:36 AM	06-Dec-2012 09:50:05 AM	06-Dec-2012 09:50:18 AM		
TAA_DS_01	14.1.1	Validate Statistics	Completed (Pass)	pnguyen	08-Oct-2012 09:37:36 AM	19-Oct-2012 10:53:05 AM	19-Oct-2012 03:52:05 PM		
TAA_DS_01	14.1.1	QC	Superseded		08-Oct-2012 09:37:36 AM				
TAA_DS_01	14.1.1	Revise Program(1)	Completed	kreece	21-Nov-2012 12:39:13 PM	04-Dec-2012 09:42:05 AM	04-Dec-2012 09:42:05 AM		Update so that all subjects in the ‘Not Randomized’ column show as discontinued. For all columns, calculate reason for discon % using discontinued as denominator
TAA_DS_01	14.1.1	Verify Revision(1)	Completed (Pass)	pnguyen	21-Nov-2012 12:39:13 PM	06-Dec-2012 09:43:05 AM	06-Dec-2012 09:43:05 AM		Update so that all subjects in the ‘Not Randomized’ column show as discontinued. For all columns, calculate reason for discon % using discontinued as denominator
TAA_DS_01	14.1.1	Validate Cosmetics(1)	Completed (Pass)	rwoolson	21-Nov-2012 12:39:13 PM	06-Dec-2012 09:50:05 AM	06-Dec-2012 09:50:18 AM		Update so that all subjects in the ‘Not Randomized’ column show as discontinued. For all columns, calculate reason for discon % using discontinued as denominator
TAA_DS_01	14.1.1	Validate Statistics(1)	Completed (Pass)	pnguyen	21-Nov-2012 12:39:13 PM	06-Dec-2012 09:43:08 AM	06-Dec-2012 09:43:08 AM		Update so that all subjects in the ‘Not Randomized’ column show as discontinued. For all columns, calculate reason for discon % using discontinued as denominator
TAA_DS_01	14.1.1	QC(1)	Completed (Pass)	rwoolson	21-Nov-2012 12:39:13 PM	06-Dec-2012 09:50:28 AM	06-Dec-2012 09:50:36 AM		Update so that all subjects in the ‘Not Randomized’ column show as discontinued. For all columns, calculate reason for discon % using discontinued as denominator

CONCLUSION

The rapid, reliable, and cost-effective production of FDA deliverables is the Holy Grail of pharmaceutical companies and Contract Research Organizations (CROs). Difficult to achieve in a “calm” environment, they become even more problematic as statisticians, programmers, and project managers edge closer to the submission date. The potential for miscommunication within the project team increases, seemingly exponentially, thus making critical the need for robust programming tools and practices.

An earlier paper by the author (see “References,” below) stated that well-constructed metadata and metadata access tools can significantly improve the creation of the datasets that comprise an electronic submission. Several years later, we believe this also applies to producing analysis and displays: metadata-driven applications and utilities can be integrated into standard business processes, speeding the production and improving the quality of deliverables.

Pharmaceutical projects are typically complex, involving organizing hundreds of files in wide variety of formats. Furthermore, as the submissions (or delivery) date nears, there is an increased need for quick production of hundreds of displays without sacrificing quality. Moving to a metadata-driven system at Rho has led to more efficient processes and has improved the overall quality of deliverables.

By utilizing a metadata based system:

1) Specifications for datasets as well as displays are now held in a database instead of in a document. This has led to considerable savings in the time to produce datasets and displays and has led to higher quality datasets and displays. Furthermore, the data comprising these specifications can be repurposed downstream when producing define files and to track and manage dataset and display production.

2) Once the dataset and display specifications have been entered, standard push button programs can be created to produce submission deliverables such as displays, datasets and the Define file.

3) Quality control can be automated and made more robust. Also, by using metadata as input to programs, programmer error can be decreased significantly.

4) Metadata created at the early stages of a project can be utilized throughout the life cycle of a project, thereby increasing efficiency and decreasing the amount of time and resources needed for project deliverables. For example, project level metadata created at study setup time can be used at all subsequent stages of a project.

5) We can have happier programmers. Repetitive hard-coding of title and footnote text is an unchallenging, error-prone task, especially when a deadline is looming. The metadata and the tools to access it allow programmers to focus on program functionality rather than what are, in essence, clerical issues.

REFERENCES

(1) Abolafia, Jeff and Frank DiIorio, "Brave New World: How to Adapt to the CDISC Statistical Computing Environment". 2011. Proceedings of the Pharmaceutical SAS Users Group Conference.

(2) "Updated SDTM/ADaM pilot Package".2013. Online at www.cdisc.org/sdtm-adam-pilot-project

(3) Abolafia, Jeff and Frank DiIorio,"Managing The Change And Growth Of A Metadata-Based System". 2008. Proceedings of the SAS Global Forum.

(4) DiIorio, Frank and Jeff Abolafia, "From CRF Data to Define.XML: Going "End to End with Metadata". 2007. Proceedings of the Pharmaceutical SAS Users Group Conference.

(5)DiIorio, Frank and Jeff Abolafia, "The Design and Use of Metadata: Part Fine Art, Part Black Art". 2006. Proceedings of the Thirty-first Annual SAS® Users Group International Conference.

ACKNOWLEDGMENTS

I would like to thank my colleague Frank DiIorio for his valuable input during preparation of this paper.

CONTACT INFORMATION

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

Jeffrey Abolafia
Rho, Inc.
Jeff_abolafia@rhoworld.com