PhUSE US Connect 2018

Paper CT10

Row, Row, Row of zeroes: Their significance in summary tables and subgroup analyses

Sridhar Vijendra, Ephicacy, Bangalore, India
Sumanth Horabail, Ephicacy, Bangalore, India

ABSTRACT

Every clinical trial summary table is a window to the wealth of information hidden in a trial. It is important for a statistical programmer to connect every programming task with the bigger picture of the trial objectives. Often, decisions made during programming are taken based on the SAP or protocol, which programmers are prompt to quote from. Although it is imperative to verify every decision with the study statistician, it is important for a programmer to make informed judgements on certain aspects when the information is not available in the SAP or protocol. One such aspect is that of adding zero rows to tables when the data is not available. Do you need to add that row of zeroes or do you simply leave the entire section out? This paper will delve into why this decision is not so easy and attempt to provide a guideline for programmers to use.

INTRODUCTION

Without quantifying the concept of nothing, the concept of something-more-than-nothing is hard to digest and explain. The presence or absence of zeroes and their location in a number can change the value and meaning of the number. The significance of zeroes to the numbers we use daily is extendable to the summary tables that we create in our programming activities. The presence or absence of zeroes in a summary table can make a world of difference to the information presented about the study to the medical writer, clinician and regulatory authorities.

For a statistical programmer wading through thick code, summary tables quickly become a programming job. Much has been written about the importance of understanding the data and looking at the summary table keeping in mind the trial objectives, but a programmer can easily sidestep the importance of the numbers while focusing on the format and alignments. Programming decisions are usually made based on the study documents such as the mock shells and the statistical analysis plan but there are situations when there is no information present in these documents and the programmer must consult with the study statistician or medical writer.

One such common scenario is that of adding zero rows to summary tables when the data for a row or category of rows is not available. This scenario is often seen in the early stages of the trial or in tables with many subgroups that effectively reduces the number of subjects in each subgroup. What seems like a simple programming decision shows an inherent lack of understanding about the information that is presented in those zero rows. The more important and relevant question is – are those zero rows saying something important? Will leaving them out hide information about the trial for the statistician, medical writer or the clinician? As a routine, programmers reach out to statisticians to get an answer to this common question. In most CROs, study statisticians are not across the hall or even an instant message away. With most study teams globally distributed, an answer from the statistician may be at least 12 hours away. With a deadline looming, should the programmer add those zero rows or leave them out? Is it possible for the programmer to make such a decision?

Adding the zero rows or leaving them out may be neither a primary issue nor a critical one, but it has the potential to save time and re-work in the event of tight timelines where last minute changes are frowned upon by the programming team. The fact these changes are usually required to be implemented when the statistician's comments are received, and time is at a premium, makes it that much more frustrating. With no widely accepted programming flowcharts or standards for SAS® programs that generate TLFs and analysis data sets, TLF programs do not easily allow for quick changes. And obviously, the more complex the table format, the more intricate the code that creates it. In addition, given the transient nature of programming teams, the programmer making the last-minute change is not usually the one who wrote the program in the first place, which can cause further delay.

Tables that present summaries by subgroups such as gender, race, previous history and other subgroups relevant to a study also present significant questions to the programmer regarding zero rows, categories with no subjects and so on. And there is no easy answer for these questions. Discussions on this topic with study statisticians and requests for tips to solve this mystery will usually result in a thumb rule that a programmer can apply. While thumb rules do work, they are usually too broad to handle study-specific exceptions. The study statistician is usually duty-bound to pass on the question to the medical writer who has a greater interest in the numbers since they are the ones to interpret it in text for the study report.

This paper aims to bring the attention of statistical programmers to this very subtle aspect of zero rows in summary tables by exploring the common scenarios containing rows of zeroes that often present themselves in everyday summary tables. Through discussions with a statistician, an attempt is made to identify what lies behind this decision about the inclusion or exclusion of zero rows. The paper concludes with a summary of the observations

from the sample scenarios and recommendations for programmers to follow for handling zero rows in summary tables.

A THUMB RULE

The most common rule of thumb provided for this decision is very straight-forward – in case there is no data available for a row, add zero rows for tables related to the SDTM intervention class (Exposure, Concomitant Medications etc.) and findings class (Labs, Vitals, ECG, etc.) but leave them out for the event class (Adverse Events, Medical History, etc.). At a quick glance, this thumb rule appears to work since we rarely fill out dummy zero rows for the typical adverse event tables such as those by System Organ Class and Preferred Term as shown in **Figure 1**. Conversely, tables involving exposure such as the one shown in **Figure 2** almost always involve a duration category that is filled with zeroes for the categories for which data is not available. So far, so good.

Figure 1: A typical AE summary table which follows the thumb rule

Incidence of TEAEs by System Organ Class and Preferred Term

System Organ Class/ Preferred Term	Treatment A (N=xxx) n (%)	Treatment B (N=xxx) n (%)	Total (N=xxx) n (%)
Number of subjects with at least one TEAE	xx (xx.x%)	xx (xx.x%)	xx (xx.x%)
System Organ Class A	xx (xx.x%)	xx (xx.x%)	xx (xx.x%)
Preferred Term A1	xx (xx.x%)	xx (xx.x%)	xx (xx.x%)
Preferred Term A2	xx (xx.x%)	xx (xx.x%)	xx (xx.x%)
Preferred Term A3	xx (xx.x%)	xx (xx.x%)	xx (xx.x%)

Figure 2: A typical exposure summary table showing distinct duration categories

Summary of Exposure

	Total (N=111)
Exposure (days)	
n	110
Mean	15.3
SD	1.41
Median	15.0
Min	10
Max	26
Range of Exposure (days)	
n	111
< 7	0
>= 7	110 (>99%)
Missing	1 (<1%)

ANALYSIS OF THE THUMB RULE

As we analyzed further, more gaps propped up in the thumb rule. How does one answer the same question for special purpose domains such as demography tables? What about efficacy tables? We had not heard anything about a rule for such tables. Does the simple rule extend to summary tables with subgroups as well? As we examined more scenarios, it became clear that there were a number of factors that influenced this decision, not

limited to the medical writer's desire to see explicit zeroes in specific rows, the ability of zeroes to help identify outliers in the data and the ability of zeroes to help in ensuring a total not greater than the sum of all the subjects in all categories of a treatment arm. These factors, either alone or in combination influence the statistician's decision on the row of zeroes that is conveyed to the statistical programmer.

STUDY CONDUCT INDICATOR

One of the most important indicators of study conduct is the disposition table that presents information about treatment discontinuations and patients withdrawn from the study due to various reasons such as adverse events and protocol deviations as shown in the example in **Figure 3**. In most cases, the statistician will ask for zeroes to be added to this table for all rows shown in the CRF and sometimes, even for all rows in the "other" section of the CRF. Although this is in clear exception to the thumb rule, common sense indicates that the information presented in this table is a veritable snapshot of the conduct of the study and is invaluable in identifying inconsistencies and extreme deviations in the safety of the drug. The zero rows provide an explicit picture of the lack of deviations or unacceptable events in the study to the medical writer who words the text of the study report.

Figure 3: A typical disposition table

Subject Disposition and Withdrawals

Number of Subjects		Treatment A	Treatment B	Total
Enrolled	N	4	28	32
Premature Discontinuation from Study Treatment	n (%)	5 (100.0)	27 (96.4)	32 (96.9)
Reasons for Discontinuation from Study Treatment				
Radiographic disease progression	n (%)	0	16 (57.1)	16 (50.0)
Clinical disease progression	n (%)	0	3 (10.7)	3 (9.4)
Unacceptable toxicity	n (%)	0	0	0
Withdrew consent	n (%)	0	0	0
Death	n (%)	0	1 (3.6)	1 (3.1)
Pregnancy	n (%)	0	0	0
Discontinued by investigator or the sponsor	n (%)	1 (25.0)	0	1 (3.1)
Significant protocol deviation	n (%)	0	0	0
Lost to follow-up	n (%)	1 (25.0)	0	1 (3.1)
Adverse event	n (%)	1 (25.0)	2 (7.1)	3 (9.4)
Other		2 (50.0)	5 (17.9)	7 (21.9)

CROSS-VERIFICATION OF THE TOTAL

Laboratory shift tables as shown in **Figure 4** are a good example of tables that are error-prone and notorious not only for their layout but also for the number of stat comments they usually trigger. If the programmer follows the thumb rule verbatim, the zero rows should be in the report. The zero rows that a statistical programmer may occasionally leave out in a shift table have an important role to play. The zero rows convey to the statistician that the data for all the rows and columns indeed add up and nothing has been inadvertently missed out during the programming of the analysis data set or the table itself. As seen in the sample shift table, the low and missing rows contain all zeroes and these help to cross-verify the total seen in the bottom row and right-most column respectively for each visit.

PRESENTING KEY INFORMATION

The demographics table, as seen in **Figure 5** is key in every study and is one of the first ones to be presented in a study report since it displays the actual baseline demographics of the study sample. Aptly categorized as a special purpose domain in SDTM, it goes without saying that this is out of scope of the afore-mentioned thumb rule. However, it is important to remember the significance of the information presented in the demographics table as seen from a clinician's or regulatory authority's perspective. It is a standard practice to present zero rows for categories present in the CRF that do not exist in the data as of database lock. This information helps to corroborate or invalidate the key objectives of the drug trial regarding the target population as detailed in the protocol. Whether it helps corroborate or invalidate, it is very important not to hide this information about the study demographic.

Figure 4: A typical lab shift table

Shift from Baseline in Hematology Laboratory Parameter

Treatment Parameter (Unit)	Visit	Timepoint	Baseline level	Low n	Normal n	High n	Missing n	Total n
ABC xx mg (N = x)								
Hemoglobin [g/dl]	Day 1	2H	Low	0	0	0	0	0
			Normal	3	70	1	3	77
			High	0	0	0	4	4
			Missing	0	0	0	0	0
			Total	2	70	1	7	81
	Day 2	24H	Low	0	0	1	0	1
			Normal	0	46	4	30	80
			High	0	0	0	0	0
			Missing	0	0	0	0	0
			Total	0	46	5	30	81

Figure 5: A standard demographics table

Summary of Demographic Characteristics at Baseline

Characteristics	Statistic	Treatment A (N=xxx) n (%)	Treatment B (N=xxx) n (%)	Total (N=xxx) n (%)
Sex				
Male	n (%)	7 (53.8)	2 (33.3)	9 (47.4)
Female	n (%)	6 (46.2)	4 (66.7)	10 (52.6)
Ethnic Origin				
White	n (%)	7 (53.8)	4 (66.7)	11(57.9)
Black	n (%)	6 (46.2)	1 (16.7)	7 (36.8)
Hispanic	n (%)	0	1 (16.7)	1 (5.3)
Other	n (%)	0	0	0
Age (Years)	n	13	6	19
	Mean	13.2	13.7	13.3
	SD	1.5	1.6	1.5
	Median	13.0	13.5	13
	Min, Max	11, 15	12, 16	11, 16
Age group				
10 and Under	n (%)	0	0	0
Pre-teen	n (%)	5 (38.5)	2 (33.3)	7 (36.8)
Teen	n (%)	8 (61.5)	4 (66.7)	12 (63.2)

DATA OUTLIER IDENTIFIER

Although considered primarily a data management task, identifying outliers in the data caused possibly by incorrect data entry or by actual data points that call for more scrutiny, is the duty of every member of the study team. It is important for each of us to keep our eyes open to catch such outliers and direct queries about it to the relevant team members. It is well known that summary tables can also help us achieve the same goal, be it with zeroes or non-zero numbers. And statisticians are skilled at identifying data issues from the table without even glancing at the data set. A simple example is the Body Mass Index (BMI) section of the demographics table where the categories of 15-20, 20-25 and 25-30 are displayed. If the study inclusion exclusion criteria require BMI to be in the range of 15-25, a statistician expects to see a 0 in the 25-30 category. If there are any outliers that show up in this row are non-zero values, the statistician makes a note of such subjects to ensure they are followed up through the study.

THE EFFICACY ANGLE

Efficacy data, not covered by the generic thumb rule are better handled on a case-by-case basis. Efficacy data dealing with critical survival information and those involving primary endpoints are handled differently as compared to efficacy data that is less critical to the study. An example of an efficacy summary table collating tumor size information is shown in **Figure 6**. The zero rows shown highlighted in the table are an example of the rule to follow for all efficacy tables – when in doubt, add them.

Figure 6: A Tumor-size Efficacy Summary table

Summary Statistics for Tumor Size at Week 6 and Week 12

Visit	Statistics	Treatment A (N=15)	Treatment B (N=2)
Baseline	N	15	2
	Mean	124.4	60.5
	SD	53.48	30.41
	Median	129.0	60.5
	Min, Max	43, 211	39 ,82
Week 6	N	8	0
	Mean	147.2	
	SD	57.64	
	Median	163.9	
	Min, Max	63 ,231	
Week 6 Percentage CFB	N	8	0
	Mean	-0.7	
	SD	16.43	
	Median	-1.8	
	Min, Max	-21 ,33	
Patients with imputed data at week 6	n (%)	11 (73.33%)	0
Reason for imputation			
Died prior to week 6	n (%)	1 (6.67%)	0
Symptomatic progression	n (%)	3 (20.00%)	0
Radiological progression prior to week 6	n (%)	0	0
Radiological progression	n (%)	7 (46.67%)	0

THE MEDICAL WRITER'S CHOICE

Of all the scenarios discussed, the trickiest one that is almost impossible for the statistical programmer to crack by themselves is that of an ad hoc decision taken by the statistician or medical writer to exclude zero rows due to reasons that are not in the purview of the statistical programmer. One common example of this scenario is the exclusion of zero rows for adverse event tables by severity or in multi-parameter categorical endpoint tables that flow for many pages, simply because the rows and rows of zeroes add too many redundant pages to the report and can be excluded without any loss of information in that specific study. Since this is mostly subject to the discretion of the statistician and/or medical writer, the thumb rule simply does not apply here.

SUBGROUP ANALYSES

Tables by subgroups present several questions on zero rows to the programmer in addition to the question that has already been discussed above. The original question is particularly relevant since subgroup analyses almost always result in one or two subgroups with less than 10 subjects each, of which fewer will contain data relevant to the table. Some of the other questions that arise in summary tables by subgroups are:

- Should zero rows be displayed when there is no data present for a specific subgroup while the subgroup has non-zero subjects?
- Should zero rows be displayed when there are no subjects in a subgroup presented in the CRF?

The most common of answers to these questions are the statisticians' request for the message "No subjects meeting the criteria" to be displayed for sections that have no data for the subgroup. In case there are no subjects in the subgroup, the entire page for that subgroup is left out of the report.

WHEN IN DOUBT

Barring all the multiples scenarios, what happens when the statistical programmer is faced with a new situation where none of the above rationale apply? When in doubt, add the zeroes anyway. There is a higher chance of being correct than not. But make sure the program is easily modifiable for removal of the zeroes.

CONCLUSION

This paper addresses the not-so-uncommon issue of zero rows in summary tables. With no data available, should a programmer add those zero rows or leave the entire row out of the table? Almost every programmer can attest to asking this question at one point or the other to the study statistician. Precious time and effort in the last few days of a deliverable can be saved if the programming team is able to make a smart decision for the summary tables usually affected by this issue. This preemptive decision will be confirmed or rejected by the statistician during the review but if a programmer can get the decision right more times than not, the purpose is served in reducing re-work. The added benefit is that the statistical programmer has a keener insight into the numbers they are programming.

We introduced a common thumb rule that is suggested by some statisticians to make this decision. This thumb rule suggests the display of zero rows for tables that involve the SDTM intervention and findings class and the skipping the zero rows in tables that involve the event class. This thumb rule works in many cases, not limited to adverse events by SOC and preferred term, exposure summaries, lab summaries and the like. There are reasons behind the thumb rule and it is important to apply it with a grain of salt. Instead of a blind application of the rule, it is helpful to understand the rationale behind it and identify cases where the thumb rule can be discarded in favor of a more logical answer.

We presented various common scenarios and the rationale behind each of them. Disposition summaries present key study conduct information and the zero rows that are usually displayed are invaluable in explicitly indicating the lack of deviations from the protocol, despite this being an exception to the rule. The reason lab shift summaries follow the rule is because the zero rows help the statistician and reviewer cross-check the total in all rows and columns. Special purpose domains such as demographics need to display the information for all categories of the demography CRF irrespective of whether any subjects are present in them or not. This is necessary in order to present the full picture of the study demographic to the reviewers. Efficacy tables are a different game altogether and the general suggestion in such tables is to populate zeroes when in doubt rather than leave the rows out. Tables presented by subgroup are best handled with an explicit message stating the lack of subjects meeting the criteria for a particular subgroup. The final type that was discussed covers scenarios where all rules fall apart because the statistician or the medical writer decide to remove empty rows of zeroes for reasons that are not apparent to the statistical programmer such as reduction of report size.

In summary, a statistical programmer can make a sound decision about adding zero rows or leaving them out based on some of the reasons explained. A meaningful addition to the thumb rule will be – **When in doubt, add them anyway**. It is also very important for programmers to keep these zero rows easily removable in the program so that such changes can be done with a single macro variable update or something similar. Of course, all decisions are verified with the statistician at the time of review but having quick discussions with the medical writer and statistician prior to programming and looking beyond the numbers of the summary tables can help the statistical programmer produce outputs that are both accurate and closer to the actual report the statistician and medical writer desire to present in the study report.

ACKNOWLEDGMENTS

Thanks to the Ephicacy management for their unwavering support in making this happen.
Sridhar Vijendra would like to thank his manager Tyagrajan Swaminathan for his constant encouragement and continued support.

CONTACT INFORMATION

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

Sridhar Vijendra
Ephicacy Lifescience Analytics
Bangalore, India
sridhar.vijendra@ephicacy.in

Sumanth Horabail
Ephicacy Lifescience Analytics
Bangalore, India
sumanth.horabail@ephicacy.in