PharmaSUG 2013 - Paper IB05

The Value of an Advanced Degree in Statistics as a Clinical Statistical SAS Programmer

Mark Matthews, inVentiv Health Clinical, Indianapolis, IN
Ying (Evelyn) Guo, PAREXEL International, Waltham, MA

ABSTRACT

Clinical statistical programmers often leverage the SAS software to process data in the form of a data set, table, figure or listing. Frequently there are a set of rules, also known as programming specifications, which can enable a non-statistician to compute some of the most complex statistical results. These specifications are generally created by a seasoned biostatistician. However, with current industry trends, more advanced statistical methods are available. The statistician needs to be free from creating those programming specification and rules to explore better analysis methods. A SAS programmer with an advanced degree in statistics is quite capable of closing those gaps which makes the SAS programmer, who is a statistician, a very effective role in the clinical field.

ABOUT THE AUTHORS

Both authors have been in the clinical field for a combination of 20 years. They have spent time in both pharmaceutical and CRO companies. They both have a master's degree in statistics from universities located in Indiana and have worked directly with each other in both CRO and sponsor company. Their field of expertise has focused on the statistical SAS programming aspects of clinical trials. Mark, currently a director, has engaged in the management space over the past decade and Evelyn, currently a principal programmer is in the technical leadership space. The following are actual experiences from the authors or indirect experiences while networking with our peers while sharing experiences through project work, mentorship, and lunchroom casual business discussions. The views and expressions are of those of the authors and may not reflect others' opinions.

OVERVIEW OF A CLINICAL STATISTICAL SAS PROGRAMMER

Times change and so do the industry trends. Over the recent years, new situations have become more prevalent to statistical SAS programmers which require continuous education on such developments. Examples include the implementation of CDISC standards, newer SAS versions and platforms, advanced therapeutics, and novel designs such adaptive designs and Bayesian analysis methods. In addition, along with increasing need of data services and cost saving strategies, some of the statistician's work shifts to the programmer. For example, the senior programmer may need to produce the specifications for an analysis dataset, and the programming plan for the tables, figures, and listings. Regardless of what is being programmed, the SAS programmer will follow the detailed instructions in the specifications documents. Thus, there needs to be a set of instructions or rules that is easily understood for all the analysis. In particular, for the complicated statistical analysis, the statistician can be frequently in the situation of the extent of even placing SAS code in the instructions. This can divert needed attention away from the statistical analysis plan (SAP) which otherwise can better serve the analysis's purpose. An improved model would be for the statistician to put full attention toward the SAP as well as other areas where novel statistical methods are deployed, and have the statistical programmer, who can thoroughly understand the SAP and protocol write the programming requirements and ultimately deliver the intended results. Although an advanced degree in statistics is not required to adequately provide this responsibility, it can however, benefit the overall workflow.

ADDS VALUE TO THE PROJECT

Two types of statistics that are heavily used in clinical analysis are descriptive and inferential statistics. Descriptive statistics are simple by nature and can easily be categorized as *describing the sample*. Calculating and using descriptive statistics may not require advanced analytical skill to calculate them; it only requires the ability to program SAS. The selection of the statistics is done at the SAP and then there are programming requirements that ultimately describe how to place the statistics in a table, figure or listing. Inferential statistics on the other hand can be categorized as *drawing inferences about the population* of interest based on the sample. This too may not require

advanced analytical skill to calculate them; again, there is SAS with all its functions that do this for the programmer and there is a qualified statistician who develops and is accountable for all the analysis.

In the following three examples, although advanced knowledge of statistics may not be truly needed from the programmer, it can be shown that there is value added when that knowledge and specialty exists.

EXAMPLE 1: DETECT ISSUES EARLY

Consider the CRF design. Normally the senior SAS programmer will be involved in the CRF design or selection process, and also will review the SAP in order to ensure that the data structure and statistical design are doable using SAS. An experienced SAS programmer will easily identify key elements in CRF which crucial for SAS programming. However, they may not be able to detect the cases in which the data may not be sufficient for statistical analysis. For example, the SAP requires an analysis that can detect the relationship between Y and X, and would also want to have multiple factors, and multiple interactions. For an analysis like this, there would normally need a large sample size in order to converge, otherwise, SAS/STAT won't work due to lack of degrees of freedom. An experienced SAS programmer may not know that there can be a lack of sample size until the actual programming has been developed. After the problem has been detected, the programmer will most likely consult with the statistician, and make changes in the SAP, specifications, and the programming plan. This will result a lot of rework. However, a SAS programmer with advanced degree and highly experienced in this analysis can easily identify *upfront* that the sample size may be too small for the analysis, recommend what can be done to mitigate this risk, and ultimately can avoid a lot of rework

EXAMPLE 2: NOTHING WAS INCORRECT

A straightforward baseline to endpoint analysis on a primary endpoint with the use of a general linear model is quite a common task for an experienced statistical programmer. Most times, the table requirement would include all the proper SAS commands. Examples include the use of, say, PROC GLM that specifically spells out the CLASS variables and MODEL statement so that it is clear on the use of independent variables and covariates, etc. A seasoned statistician could easily produce the requirements and so could the programmer if experienced enough to understand the analysis and how that particular table fits within the overall analysis plan. However, consider what happens if the analysis methods needs to change. An example to share is the statistical programmer programs the table, and an independent programmer replicates the results since it is an important primary endpoint. The programmers run through their validation routine, checking all the log files etc, and it ends up that everything checks out. However, the statistician continues to assess the analyses across the board and then decides it is more appropriate to 'center the baseline' value. That is, the statistician instructs the programmer to take all individual baseline values and subtract that from the overall mean and use that resulting new variable as the baseline. Again, an experienced programmer would be able to quickly update the specifications and program the resulting analysis. Also an experienced programmer would again check the log file and look at the plausibility of the new result to make sure everything checks out. In this case, prior to the release in production, the programmer observed that the p-value did not change; it was identical to its value before the new centered baseline variable was in the linear model. The programmer then continues to redo all the steps involved in creating the new analysis and even consults with a fellow programmer (not the validation programmer of course). Every step involved with the analysis seems to check out okay and the programmer decides to present the analysis for validation; with a degree of caution because the p-value did not change even though multiple thorough checks were consistent with the specifications. Now on the other hand, if the programmer had the inferential statistical training as an advanced level, they would know immediately that a linear transformation on a linear model using Type III sum of squares will not change the p-value. It only changes the assumptions around the analysis. The result would be confidence in the analysis as well as time not lost by repetitively checking and consulting with others whether or not the programming and analysis were done correctly.

EXAMPLE 3: DETECT ISSUES LATE

Another example to share is an inferential analysis where the F-statistic was printed in the table. In this case, two similar studies were recently done prior and naturally there was a lot of "copy over" analysis, requirements, and code. The resulting F-statistic in this case came out to be in the 700s. It is of course mathematically possible to have an F-statistic that high, but the programmer, who was a statistician was prompted to look into it further because of the relatively large value when compared by the 2 prior studies. It turned out that the CRF (case report form) page of the collected data inadvertently had a different version where the value collected was the number of doses, and not the total milligrams of dose taken. However, the manual annotation of that value retained the prior study settings and the number of doses was coded as milligrams (mg). The differences between the dose counts and the total dose in mg collected were not large enough to immediately observe by individual inspection, but were detected in the use of inferential statistics Prior to the source code fix, the programming was not wrong. It met all the specifications. However, the analysis was not correct.

In the above three examples, a seasoned statistical programmer with advanced degree or knowledge of inferential statistics would perhaps be skilled to know when to investigate further and when not to. This ultimately results as

value is added to the projects with time saved and a minimization of risks of improper analysis. The applications of statistical analysis can be very thorough and done efficiently if appropriate applied statistical knowledge is present, either by formal education or experience, in those doing the programming. It may seem that no matter how detailed the programming specifications are, there can exist some level of ambiguity in the requirements. Either way, when inferential statistical methods are deployed, there needs to be a specialist that has the proper training and experience to implement such principles on all projects.

ADDS MARKET VALUE TO THE INDIVIDUAL

Transferrable skills between a statistician and a statistical programmer can enable a degree of resource flexibility in certain cases. In this example, the biostatistician fully develops the statistical analysis plan (SAP) and provides the table shells with certain analyses requirements with SAS code. Then, the programmer prepares to design the overall programming environment and then decides the best code to (re)use or develop. Somewhere along the line, there will be areas where the requirements in the TFL shell are not enough to complete or fully describe the analysis. Additionally, the source SAP may not answer the uncertainty in the analysis. A simple example would be that the title of the table shell suggests using the investigator site as a covariate in the model, but the shell requirements do not mention anything of using a covariate. So which is correct, the title or the table requirements? When this happens, it requires the programmer to approach the biostatistician so that the ambiguity can be resolved. For instances when the programmer is a seasoned statistician, the role delineation model could change such that the biostatistician would fully develop the SAP and the programmer would fully develop and own the TFL shells and furthermore the programming and analysis of the resulting TFLs. In phase III clinical trials where the studies are sometimes large and complex, this flexibility of role delineation has proven valuable to best fit the overall environment. The biostatistician in this case can perhaps now spend time on more challenging problems and expand their role in consulting and teaching others the fundamentals of what they do and can do. And of course the programmer owns more of the clinical data flow which eliminates another layer of communication. Furthermore, the programmer can now use the situation to find ways to improve the programming design of the TFL cycle such as dynamically linking the requirements with the actual coding programmatically for a more complete and automated system.

Consider the environment of Phase I trials where the data is much different than later phases and thus posing different analysis and validation requirements. For example, it is common for the development of the TFLs and data sets not to require an independent validation program to replicate the results. Sometimes you can verify the results from inspection in cases where there are few enough data points to do so. Moreover, there can be phase I trials where there is no programming role; the biostatistician does all the programming. Or, a programmer who is a seasoned statistician who will program the data sets and TFLs will in fact create and own the statistical analysis plan. It is these types of transferrable skills that allow flexibility in the work environment in clinical analysis and can add to the value of the individual by having expertise in a diverse amount of tasks.

The authors have experienced cases where a programmer, who is a statistician can effectively use their collaborative skill set in order to take on more responsibilities than would a programmer without an advanced degree in statistics or equivalent experience. A trend in the industry that they have seen is that some statistician responsibilities are shifting to the programmer: such as the table specifications, mock shells and other parts of the analysis plan. These tasks require the appropriate amount of experience. A programmer with a statistics degree can reduce the learning curve and sometimes add quality to the overall result when the statistical methods are complex. Having transferrable skills in this environment does provide career growth opportunities for the programmer.

The above example discussed situational role delineation. How about more permanent and long term situations than adapting your tasks to the work environment? It is very possible for a statistician who is a very good programmer to evolve or change their career into another specialty. For example, an individual who is titled “programmer” could spend a few years in the programming TFLs and data sets, and then switch gears to permanently move roles and now be titled a “biostatistician” and spend years there developing SAPs and contributing to the overall therapeutic area developments. In fact, at the time of this publication for the past 4 years, one of the authors have supported 4 programmers, who had masters degrees in statistics and the motivation to do biostatistics, to permanently move into the role and title of a biostatistician. That has been an average of 1 per year lately.

There is a similarity in what a statistical programmer and a biostatistician does. But what about other functions where there is less similarity. In a case where a programmer with an advanced degree in statistics has spent time in data management developing SDTM domains and Case Report Tabulations (CRT), although there is a need for CDISC and SAS programming expertise, there had yet to be a need of applied statistics in that role. With some career paths, it is advantageous to experience different clinical functions and expose themselves to the larger part of the clinical drug cycle. It is perhaps easier for a programmer who is a statistician to move into other functions such as data management rather than the opposite; a data manager who is not a statistician to move into the biostatistics role.

However, there have been many cases where the data manager and statistical programmers can leverage their talent and more easily change roles; especially when both roles have a set of collaborative or common skill sets.

Being able to move from one organization to another is valuable to the career if one is interested in the breadth of experience. It can also be beneficial to your own net worth; for example, about a decade back, one of the authors moved from a SAS programming Information Technology department to a Statistical Sciences department and took on a SAS/statistical analyst role. With much surprise, the author soon thereafter had a meeting with their supervisor and was explained the pay increase that will be given solely for having a masters degree in statistics. It was assessed that the market value was higher than in a position of statistical analysis than in the information technology department as a systems analyst. In some cases, and depending on your environment and company, an individual who has a master's degree in statistics can have a competitive advantage in terms of compensation.

The SAS programmer who is a statistician can be in a position to increase the flexibility of operating across roles and functions. This skill set ultimately enables fewer limitations on career opportunities and can add value to their career by being more versatile and effective in multiple functions.

ADDS MARKET VALUE TO THE EMPLOYER

An advanced degree in statistics can also add value to the employer as well in both pharmaceutical companies as well as CROs. The core of a sponsor organization is research and development of medicines with an emphasis of selling drugs in order to reinvest in developing even more. The core of the CRO is to conduct any essential work on behalf of the sponsor company for a fee for service with an emphasis of performing the work cost efficiently so that the symbiotic relationship can sustain itself[1].

In a prior example it was suggested that a statisticians can enable themselves to spend more time designing trials with less hands-on with the analysis. This can be considered a benefit to a sponsor company. The statistician in this case can now place their time into new, unmet areas of research. The specialization of applied statistics can be an integral part of the core of a sponsor company. Likewise, the CRO can move some of the statistician responsibility to the programmer which can yield higher profitability or cost savings depending on the resource models.

With the CRO, there is a need to continuously seek business by attracting sponsor companies for continuous and new business. The CRO responds to the sponsor company when solicited and provides the best statements and facts that would perhaps entice the sponsor to use the CRO services. An example would be for the CRO to state that, say, 25% of all statistical programmers have a master's degree in statistics. And also, during the actual bid defenses, there can be a single individual who is representing not only statistical programming, but biostatistics as well which makes the resources needed for the defense to be lean and more efficient. In the case where a programmer, who is a statistician, is representing both the programming and statistics functions, is capable of providing accurate information. And in the event there is a specialty response needed that is outside the knowledge of the individual, then they know when and where to consult for direction and advice.

There are numerous articles around the use of six sigma principles in the role of statistical SAS programming. There was a situation where the fully trained master black belt, who was not a statistician, was explaining the statistical methods used to analyze the project; the "Analyze" in the DMAIC process. When the discussion of the use of ANOVA and linear regression was brought up, there was a general response by those on the team, who were not statisticians, saying that the applied statistics very complicated and therefore could not engage fully in the mechanics of that step on the six sigma process. In fact, the interest on how it works was not of main interest to the team in general; they were only interested in the outcome. At the same time, a programmer on the six sigma team, who was a statistician by education, expressed that the ANOVA and other methods were fundamentally simple and intuitive. When there are six sigma projects with statisticians being part of the core or extended team, it can result in less explanation of the mechanics of the applied statistics and more emphasis on analyzing the outcome or the other steps DMAIC steps of the six sigma process.

Resourcing and cost is important to both a pharmaceutical and CRO company. Throughout this article, there were examples where time has been saved, and rework has been minimized, or the analysis was done right the first time, as a result of transferrable or collaborative skill sets from programmers having an advanced degree in statistics. Without this benefit there would be a need for additional time, or resources to be applied to the efforts associated with the process. Also, this article suggested that the statistician role should spend more time on the novel and modern analysis types and less time on the essential work on processing and analyzing the data from the clinical trials. When more responsibility is placed on the programmer, then the industry as a whole can perhaps benefit from this by exploring new simulation techniques, or further invest in adaptive designs and Bayesian analysis. A programmer that can extend their responsibilities that at statistician traditionally does can certainly add value to an organization and perhaps even the industry as a whole.

CONCLUSION

SAS programming and applied statistics is indeed an integral part of the clinical drug development cycle. The actual responsibility for all statistical work associated with clinical trials will lie with an appropriately qualified and experienced statistician, as indicated in ICH E6. When the SAS programmer has specialized skills in applied statistics, this role can substantially extend their traditional responsibilities into the larger part of the statistical aspects of clinical trials. This can, in turn, allow for a biostatistician to free up their time so that their role can further extend into new areas of research and scientific and statistical methods. A formal degree in statistics can provide a solid foundation of such required skill. And through daily experience applying this knowledge will enable opportunity for the SAS programming role to further develop the skills needed to extend their responsibilities. When a SAS programmer is skilled in applying statistical methods, it can ultimately create value opportunities for the projects, themselves, and their employer.

Opportunity	***Category***	***Value Added Impact***
Projects	Time/Cost	Can reduce requirement discussion for complex analysis
		Can minimize rework; issues detected early or late
		Frees up time for the statistician
	Quality	Lowers risk of wrong analysis for advanced statistical methods
		Another qualified set of eyes on complex methodologies
Career	Transferrable Skills	Reduces limitations on where to work; across departments
		Adds learning experience variety throughout career
	Extended Responsibilities	Broader exposure to various parts of the business and industry
		Can earn higher title based on responsibilities
	Recruiting	Can be looked at first – “Preferred Background”
		Can possible earn a higher pay
Employer	Transferrable Skills	CRO selling point
		Better resource management
		Employee retention perk
	Extended Responsibilities	Enables the Statistician to engage in less essential work and more on advanced statistical methodologies

REFERENCES

[1] Minjoe, Sandra and Matthews, Mark (2011): “Similarities and Differences in SAS Programming Among CRO and Pharmaceutical Industries.” http://www.lexjansen.com/pharmasug/2011/IB/PharmaSUG-2011-IB05.pdf

CONTACT INFORMATION

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

Name:	Mark Matthews	Evelyn Guo
Enterprise:	inVentiv Health Clinical	PAREXEL International
Address:	4745 Haven Point Blvd	195 West St
City, State ZIP:	Indianapolis, IN 46280	Waltham, MA 02451
Work Phone:	317.773.5901	978.313.1811
E-mail:	mark.matthews@incentivhealth.com	evelyn.guo@parexel.com