Implementing CDISC Using SAS®

An End-to-End Guide

Chris Holland and Jack Shostak
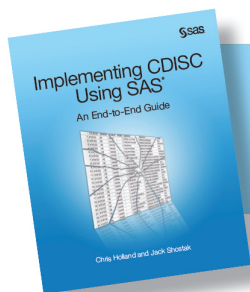
From *Implementing CDISC Using SAS®.* Full book available for purchase <u>here</u>.

Contents

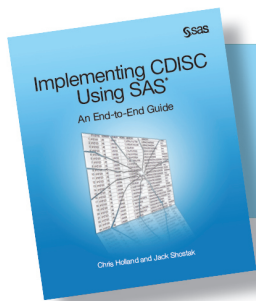# Chapter 1: Implementation Strategies

Why "CDISC It"?

The decision to adapt to CDISC standards within an organization or for a particular clinical development program is not an easy one. Many large pharmaceutical companies have their own internal standards and tools built around those standards. For these companies, the decision about if and when to switch to CDISC was, and in many cases still is, fodder for heated debates. Upper management may cringe and balk at proposals to implement CDISC standards when they see the up-front price tag associated with it. Some organizations got burned by trying to implement the Study Data Tabulation Model (SDTM) too early, before FDA reviewers were prepared for it and before the standard had stabilized to a point where it was backwards compatible. Without an FDA mandate, the rationale for taking a wait-and-see approach seemed to have its merits.

Despite these concerns, set-backs, and growing pains for CDISC, the groundswell of support continued to rise. Mergers and acquisitions have persisted throughout the pharmaceutical industry, and behind the scenes of each merger are the data managers and SAS programmers who've worked at the same desk year after year, but have seen their employer name change three to four times. Throughout it all, with a change in the employer came the change in the case report form (CRF) designs, variable names, and data formats for the different compounds on which they worked. When it came time to integrate the data for a regulatory submission, a substantial amount of time was spent deciding on the structure and variable names to be used for the integrated database. And that was just the beginning. The time spent doing the actual conversions and integration is often much greater. As the programming hours piled up, those involved started to see the merits of having a standard across the industry.

Pharmaceutical and biotech companies weren't the only organizations undergoing mergers. During the late 1990s and early 2000s, many CROs consolidated as well. In addition to the numerous data standards they had to keep track of among their various clients, CRO SAS programmers also had to deal with different data formats being used internally due to consolidation with other CROs. Some at these CROs got to work on integration projects involving compounds that, at each new phase of development, had been passed from one organization and CRO to another. As a result, even the most basic key identifier of any clinical trial dataset, the subject ID, was sometimes uniquely named within each study. So as the programming hours piled up, the key decision makers at CROs started to see the merits of having a data standard across the industry.

Yet this grass-roots initiative to develop industry-wide standards would not have gotten off the ground without the support of the biggest consumer of clinical trial data of all, the US Food and Drug Administration. Although some FDA reviewers become accustomed to data formats they see from certain sponsors, most still have to deal with completely different data formats and structures from one sponsor to the next. This might not have been so cumbersome in the days before the Prescription Drug User Fee Act (PDUFA, commonly pronounced *puh-DOO-fa*) first became effective. Before PDUFA, a review clock was non-existent and 2-year reviews of New Drug Applications (NDAs) and Biologic License Applications (BLAs) were the norm. However, with the passage of PDUFA, review cycles were originally mandated to be 12 months (and are now down to 10 months). With those review clocks, along with increasing expectations to carefully inspect the electronic data that were packaged with NDA and BLA submissions, reviewers found themselves having to do more with less.

The aftermath of some pivotal events in 2004 put even more pressure on FDA reviewers. One was the investigation of suicidality risks among children on antidepressants. The other was the withdrawal of Vioxx from the market. Because of these two high-profile safety concerns, doctors, patients, and sponsors all suddenly had a vested interest in knowing whether the drugs they were prescribing, taking, or selling to treat depression, arthritis, or any number of spin-off indications were adding on risks that outweighed the benefits. The brunt of these class-effect determinations fell on the FDA clinical and statistical reviewers who were the only ones who had access to all the data that would allow them to make the informed decisions that the public, doctors, industry, congressmen, and media were all suddenly demanding. However, when the drug class under consideration involved 10 different compounds from as many different sponsors with as many different data formats, this was no easy task. "Wouldn't it be great," some FDA reviewers asked, "if we had all the data we need in one giant database?" Fortunately, within the FDA, certain reviewers, team leaders, and division directors all started to see the merits of having a data standard across the industry. To coin a common phrase of one particular FDA division director who had a penchant for promoting data standards at industry conferences, the mantra of the late 2000s became "just CDISC-It."

Evidence of FDA's support of data standards is not only found at conference podiums. Since the draft release of version 3.1 of the SDTM Implementation Guide (IG) in October, 2003, the FDA has issued a number of documents indicating their support of data standards. A summary of these appear in Table 1.1.

Table 1.1: FDA Documents and Events in Support of Data Standards

Time	Event
July, 2004	eCTD study data specifications reference the SDTM for tabulation data
March, 2006	"Development of data standards" listed as opportunity #44 in the FDA's Critical Path opportunities list
September, 2006	SDTM/ADaM pilot project review completed and results presented at the CDISC Interchange
September, 2006	Old e-NDA guidance document is withdrawn (leaving the eCTD study data specifications as the only guidance relating to submission data)
December, 2006	Proposed rule to require electronic data with submissions is released in the Federal Register
May, 2008	First version of the PDUFA IV IT plan is released, making numerous commitments to the SDTM
October, 2009	Version 1.5 of the study data specifications released, making specific reference to the Analysis Data Model (ADaM) standard for analysis data
March, 2010	Version 1.0 of the CDER Data Standards Plan released, providing a commitment to CDISC standards
May, 2011	The "CDER Common Data Standards Issues Document," version 1.0 is released, stating that CDER is "strongly encouraging sponsors to submit data in standard form"
February, 2012	Draft FDA guidance document *Providing Regulatory Submissions in Electronic Format -- Standardized Study Data* (a.k.a. "the e-Data Guidance") released to establish "FDA's recommendation that sponsors and applicants submit study data in a standardized electronic format"

Nowadays, the legion of CDISC implementers is tangible to any attendee of a SAS user conference struggling to find an empty chair in a session that has anything to do with CDISC. Managers are preaching the data standards gospel, software vendors are demonstrating their tools that make use of CDISC data, FDA presenters are promoting their preference for CDISC, and FDA documents are recommending the SDTM and ADaM models as the format to use for sponsors' NDA and BLA submissions.

Despite all this, electronic data submissions are still (believe it or not) optional. Many compounds struggle to get beyond Phase I of development. Implementers have struggled with different interpretations of the standards and with study-specific data that don't seem to have a clear-cut home in the SDTM IG. So while the decision to "CDISC-It" for a regulatory submission has become more clear-cut, the stage when the CDISC conversion takes place, and how far to go with it, is less lucid. Study phase, organizational size, standards know-how, and the amount of resources all factor in to the decision-making equation.

Which Models to Use and Where

The advantages of having universal data standards are largely geared toward users of the data for review or analysis. Across studies, medical and statistical data reviewers and analysts, whether they are on the sponsor side of the equation or on the regulatory review side, are starting to benefit from having nearly instant familiarity with how data is organized for any given study. This holds true whether the data are non-clinical SEND data, SDTM tabulation data, or ADaM analysis files.

However, for those who are responsible for putting the data in these standardized formats, there is much more work involved. Before data standards, there was often just one set of data provided with NDA and BLA submissions. Many of these datasets tended to be a hybrid between the raw CRF data and the analysis files. The structures and variable names often matched those of the original database tables and therefore required little manipulation. Now, not only do you have to worry about a potentially labor-intensive conversion process from the raw data tables to the SDTM domains, there is also the need to then create ADaM datasets from the SDTM domains.

For organizations with plenty of resources to devote to the implementation of standards, this process might be manageable. CROs who conduct a high volume of conversions for their clients have an opportunity to streamline their implementation process with each new iteration. Certain technologically advanced organizations such as software companies with proprietary electronic data capture (EDC) systems and expert knowledge of the data standards are capable of developing automated tools to assist with the conversion process from the raw CRF data to fully compliant SDTM domains and subsequent ADaM data files.

For those organizations with high volume and adequate resources, the question isn't always whether they are capable of implementing the SDTM and ADaM, it's more a question of whether the effort is worth it given the phase of development (the earlier the phase, the less likely a drug will ever advance to submission). For everybody else, the wait-and-see approach might be more appealing, given their lack of expertise and resources.

Eventually, however, certain medical products will advance in development and when they do, it is better to be prepared. As such, the objectives of this book are to provide the following:

- Considerations for deciding when to start implementing CDISC standards
- Advice for how to get started with CDISC implementation and how to move forward with it
- Tools based on SAS software to assist with the creation of CDISC data and metadata documentation (and instructions on how to use them)
- Information on how to check CDISC data for compliance
- Information about tools for using CDISC data for analysis and review

Starting with the Clinical Data Acquisition Standards Harmonization (CDASH) Standard

The best way to adapt to being a CDISC organization is to start implementing standards at the initial step of data acquisition—the CRFs. The Clinical Data Acquisition Standards Harmonization (CDASH) (http://www.cdisc.org/cdash) standard was created in response to opportunity #45 in FDA's Critical Path Opportunities list, which was titled "Consensus on Standards for Case Report Forms." Although part of the initiative was to standardize the look and feel of CRFs, a big part of the initiative in the eyes of CDISC implementers was to standardize the variable names of data elements being captured in the clinical database. Having such a standard that was consistent with SDTM terminology would make the conversion to SDTM much easier. With a CDASH structure behind any data management system, certain SDTM domains, like adverse events (AE), demographics (DM), and concomitant medications (CM), are almost instantly SDTM-ready with the initial data extract to SAS datasets. In total, 16 domains are covered by the CDASH standard, covering those that are common to most therapeutic areas and types of clinical research. For further reading, the CDASH-published document (version 1.1 was published in January, 2011) contains implementation recommendations, best practices, regulatory references, and loads of other information pertinent to CDASH.

For any organization starting from the ground up, implementing CDASH should be an easy decision because it precludes the need to develop a new organization-specific standard. However, unless the data management system happens to come pre-packaged with CDASH default templates, implementing an existing standard can still require a lot of work. Without these templates, one important element to a successful implementation is making sure the proper know-how is put to work. Just a basic knowledge of CDASH might not be enough. Having a breadth of knowledge that spans CDASH, SDTM, and ADaM can help prevent you from, for example, having variable names in the source data that conflict with variables in the SDTM or ADaM data. A careful deployment with proper attention to downstream standards can save you from unnecessary variable renaming later on.

Larger, older, or high-volume organizations with established data management and EDC systems that were put in place long before CDASH probably wouldn't benefit as much from an immediate switch to CDASH because of the impact it would have on resources and timelines. For these organizations, a certain amount of creativity has had to go into the thoughts and ideas behind how and when to start with the CDISC implementation process.

Whatever the situation, whether the true source data is from an entirely CDASH environment or from something that resembles nothing of the sort, the source data can be considered just various shades of gray in the eyes of an SDTM implementer. Before delving into the programmatic conversion process, the very important step of mapping out a conversion plan needs to be discussed.

Implementation Plans and the Need for Governance

Before an actual CDISC implementation takes place, whether it is a conversion from CDASH to SDTM or the creation of ADaM data from the SDTM, it is often a good idea to document the precise mapping from one data source to another. The advantages of this are three-fold:

- it allows the work to be handed off from the planner(s) to programmer(s), thereby obviating the need to have these functions performed by the same individual.
- it provides a plan to the programmers that has been discussed, reviewed, and approved ahead of time. It will also prevent ad-hoc decisions by one programmer conflicting with those of another on the same project.
- it provides a specification that the final work product can be checked against and referred to along the way.

Anyone who has spent much time trying to implement a CDISC standard has probably quickly realized that, despite efforts to the contrary, much of it is subject to interpretation. Consequently, there is a strong likelihood that one person's interpretation is different from another's, and herein lies the foundation for another form of conflict relating to standards—the friction between two or more strong-minded individuals who each have their own interpretation on how to implement.

In order to handle this inevitable problem, many organizations have developed a form of governance where decisions relating to controversial issues are agreed upon by a group of experts. The process by which these issues are presented to and decisions are made by a governing board can vary. The board can either be responsible for reviewing and approving all document specifications developed within an organization or they can only get involved to weigh in on certain issues, especially the overarching ones that are likely to affect all projects.

For smaller organizations, use of a governing board might be unnecessary or impractical. Mapping decisions can either be made by senior personnel or by outside consultants. Whatever the size or status of an organization, in order to avoid conflicts later on, reviewing and approving mapping specifications before the actual work begins can, at the very least, prevent bad decisions from being made simply because they reflect what has already been done.

SDTM Considerations

As mentioned earlier, the decision about how and when to implement the SDTM is not always an easy one. Waiting until a Phase III study is unblinded and a pre-NDA meeting occurs can often mean having to convert a whole lot of data in a short amount of time. On the other hand, converting all studies, starting with the first-in-man Phase I, can mean spending a lot of effort on conversions for studies that might never even get into late Phase II or Phase III trials, where the benefits of SDTM conversions can really pay off.

Organizations struggling with these decisions should consider the following questions:

- Do you have the proper expertise and resources to implement the SDTM?

  Proper and compliant implementation is important in order to ensure that tools that depend on standards work properly and users of your data (such as regulatory reviewers) have a pleasant experience doing so. Although the objective of this book is to help make the process easier, it will not teach the subtle details of the SDTM. The best reference for that is the most recent version of the

SDTM IG. It is full of details and instructions that should not be overlooked. For trickier problems, such as how to model data that doesn't seem to have an explicit domain, seek advice from consultants or online experts on any of the various message boards available. But even with the proper expertise, the conversion process can be a tedious one. Make sure you have sufficient resources to conduct a proper implementation.

- Do you have enough studies in the pipeline that would allow for an efficient and steep learning curve if every study were to be converted?

  Like everything, practice makes perfect, and the less time you spend between implementations, the less you tend to forget how things were done the last time around. As such, a one-off SDTM conversion will not allow you to fine-tune the process with subsequent iterations.

- Do you have a stable environment to allow automation of certain parts of the conversion process from study to study?

  Foundational changes, such as corporate mergers or your EDC vendor going out of business, are hard to prepare for. In some situations, however, you might be able to anticipate the likelihood of having different database designs across studies. If the designs do change, then you'll have trouble building an automated conversion processes across studies. The first conversion will be a learning experience regardless. But with each subsequent conversion, the more similarities there are with the raw CRF data across studies, the more opportunities you will find to make the conversion more efficient, such as using SAS macros or standard programs for converting certain domains.

- Do you plan on using any tools that could make downstream processes such as data cleaning, safety reviews, or analysis more efficient when used on SDTM data?

  Certain off-the-shelf tools can make data review, particularly safety data review, easier if the data are SDTM-compliant. If you would like to produce patient profiles or other reports and summaries that review tools that leverage the SDTM can create, then you will certainly benefit from an SDTM conversion. Some of these review tools will be discussed in this book.

- What phase of development are you in?

  Many regulatory guidance documents provide advice about how to incorporate safety data into a submission. They tend to differentiate between Phase I safety data from healthy volunteers and those from Phase II and III studies that are more pertinent to the population, dose, and treatment regimen being considered for approval. You must also consider the attrition rate of experimental therapies. Products that eventually make it to a regulatory submission are the exception rather than the norm, and when or if they do make it to submission, not integrating or converting the Phase I data might be an option to consider because such data, aside from potential PK results, PD results, or both, are less relevant to a review of a product's safety and efficacy. Products at later stages of development, however, might reap better rewards as a starting point for implementing the SDTM.

- Should I consider a staged approach?

  Perhaps you or your organization lacks the resources or expertise for a full-blown SDTM conversion. You might still benefit from having certain key domains, such as adverse events, demographics, concomitant medications, and laboratory data in a format that, if not fully SDTM-compliant, is pretty close. Doing so will facilitate the development of standard programs, might be sufficient to use certain tools, and will make a full conversion, if required later on, that much easier. However, keep in mind that an FDA submission will likely require a fully compliant implementation.

ADaM Considerations

The first version of the Analysis Data Model (ADaM) model document was released in final form in December of 2004. It contained general considerations with respect to analysis datasets. Starting in April 2006, the ADaM team began working toward two significant goals:

- To define a standard data structure that would work well for many common analysis needs
- To create an ADaM Implementation Guide

Around the same time, the idea for a mock submission that FDA reviewers could use to see how well both the SDTM and ADaM data standards met their needs for a mock review started to gain some traction. This idea developed into the first SDTM/ADaM pilot project. During the course of a year, volunteers from industry worked feverishly to get this sample submission together, and volunteers from FDA worked diligently to closely evaluate the data, compile their comments, and discuss their findings. The constructive feedback assisted the ADaM team in its work on a new version of the model document and the first-ever implementation guide.

Drafts of the model document (ADaM version 2.1) and the implementation guide (ADaM IG version 1.0) were posted for public comment in May, 2008. Final versions of both documents were published in December, 2009 and serve as the basis for topics relating to ADaM in this book. They can be found on the CDISC web site at http://www.cdisc.org/adam.

The significance of this implementation guide should not be understated. Finally, a document existed that would allow industry to standardize the important analysis datasets that statistical and medical reviewers needed to verify sponsor results, assess the robustness of results, and explore trends in the data. Knowing that the format and structures of the data had, to some extent, been previously vetted among FDA reviewers gives implementers a certain degree of confidence that allows them to scrap their company or drug-specific formats for a new CDISC standard that will eventually have wide-spread familiarity.

The ADaM 2.1 model document highlights certain fundamental principles relating to ADaM data. As stated in the document, analysis datasets and their associated metadata must have the following characteristics:

- Facilitate clear and unambiguous communication
- Provide traceability between the analysis data and its source data (ultimately SDTM)
- Be readily usable by commonly available software tools

Further, analysis datasets must have the following characteristics:

- Be accompanied by metadata
- Be analysis-ready

The decision about whether or not to implement ADaM standards within an organization should be an easier and more straightforward one compared to the SDTM decision. First of all, an assumption with ADaM data is that there is corresponding SDTM data to which your ADaM data can be traced. So, the first question you have to ask is whether SDTM source data exists. If so, the next question is how extensive of a study, from an analysis perspective, are you dealing with. The effort to create ADaM data from SDTM data for small, safety- and PK-based Phase I studies should be balanced against a potentially limited benefit. This is because a) analyses for such studies are usually quite basic and b) analysis datasets from such trials are rarely expected for a regulatory submission. It is highly recommended, however, that at least the ADSL dataset be created. This one data set, which includes just one record per subject, is the minimum dataset requirement for an ADaM submission. Even for small, early-phase trials, it can be useful as a single source for capturing flags and certain key information relating to each subject's experience in the trial.

On the other hand, even with SDTM source data from a controlled clinical trial and a plan for inferential statistics, implementing ADaM might not necessarily be a foregone conclusion. One approach, although not recommended, is to instead implement what many refer to as SDTM+. With this approach, variables are added to the SDTM domains to assist with analysis. These variables can include treatment group identifiers, extra flags, conversions of SDTM dates to SAS date variables, and, perhaps, change from baseline calculations. While this approach might be convenient for some, it is not recommended for two primary reasons:

- Variables added to SDTM+ datasets are likely to vary across studies, submissions, and sponsors, and therefore do nothing to promote standards.
- SDTM+ datasets are usually ill-suited for providing the traceability necessary for complex computations and derivations, such as missing value imputations.

As such, the authors of this book recommend using ADaM whenever the following two conditions are true:

- A corresponding set of SDTM data exists.
- Non-trivial derivations, imputations, and/or inferences are calculated according to the statistical analysis plan for the given study.

Not all analysis data will fit into one of the predefined analysis data structures, such as the ADaM Basic Data Structure (BDS). If you are *not* using the specific data structures mentioned in ADaM documents, you should at least consider adhering to the basic principles for analysis datasets mentioned previously.

The conundrum faced by institutions looking to implement CDISC standards for the first time, particularly for early-phase trials, is that while analysis data of some sort is almost always needed for producing summaries for a clinical study report, and resources and timelines are based upon this assumption, the SDTM data is, in many ways, primarily needed for submissions. As mentioned in the beginning of this chapter, the "old days", when some organizations went straight from raw CRF data to analysis data, or a hybrid of the two, had that distinct advantage: one less time-consuming step to produce analysis data.

With this in mind, the authors concede one scenario in which you might choose to go directly from raw CRF data to ADaM, without an SDTM conversion in the middle. This one scenario would be when, due to short timelines and limited resources, the only other alternative is to not follow the ADaM standard at all and to not create SDTM data at all. Figure 1 displays both scenarios.

Figure 1.1: Scenarios for Implementing CDISC Standards
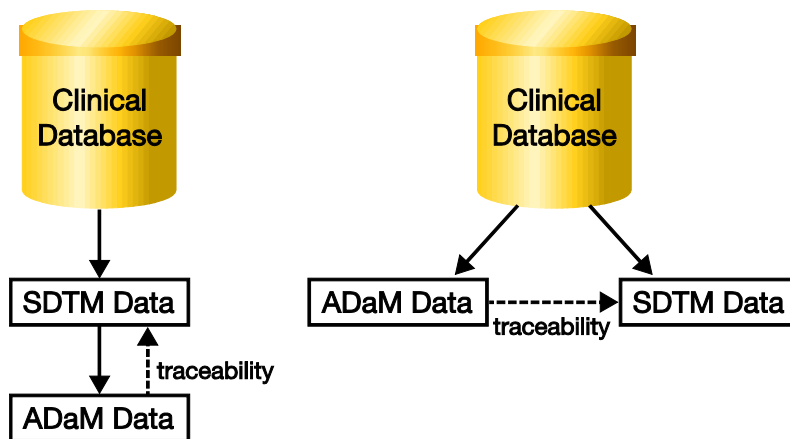
Option A (preferred): Linear conversions from the clinical database to SDTM to ADaM.

Option B (**not** recommended): SDTM is created at the same time as or after ADaM. Traceability back to SDTM is done retrospectively.

If you choose Option B, the primary risk you face is that when, or if, SDTM data are created, it could be rather difficult to ensure that records from the pre-existing ADaM datasets can be properly traced to these new SDTM data. In some cases, you might be better off creating ADaM data anew, although this

would result in duplication of work. It should also be noted that the so-called ADaM data would not truly be following the model until there were metadata and SDTM data to which they could trace back.

The authors hope that the tools based on SAS software covered in this book and the information on how to use these tools will make conversions to CDISC standards easy enough to balance out with the efficiencies gained by using CDISC data for exploration and analysis. In other words, we hope to provide readers with information and tools that make Option B an unnecessary one to consider.

Chapter Summary

The motivations for having industry-wide data standards are multi-fold. The decision about when and how to adapt to any CDISC standard is, however, a bit more complicated. In this chapter we presented some considerations to assist with these decisions. In this book, we will be providing many tools and examples to make the conversion process easier.

# Index

About The Authors

Chris Holland has been a SAS user since 1990. He currently works as a Biostatistics Director for Amgen and has prior experience heading up biostatistics departments for Sucampo Pharmaceuticals, MacroGenics, and Micromet. He has also worked as a statistical reviewer at the Center for Drug Evaluation and Research in the U.S. Food and Drug Administration. There he served as the technical lead for the SDTM/ADaM Pilot Project review team, which included FDA staff involved with assessing the suitability of CDISC standards for meeting the needs of medical and statistical reviewers. Holland continues to be active in the CDISC community, particularly with the ADaM team. He received an MS in Statistics from the University of Virginia, and a BS in Statistics from Virginia Polytechnic Institute and State University. He is an Accredited Professional Statistician™ by the American Statistical Association.

Jack Shostak, Associate Director of Statistics, manages a group of statistical programmers at the Duke Clinical Research Institute. A SAS user since 1985, he is the author of *SAS Programming in the Pharmaceutical Industry*, and coauthor of *Common Statistical Methods for Clinical Research with SAS Examples, Third Edition*, both published by SAS Press. Shostak has published papers for the Pharmaceutical SAS Users Group (PharmaSUG) and the NorthEast SAS Users Group (NESUG), and he contributed a chapter, "Reporting and SAS Tool Selection," in the book *Reporting from the Field*. Jack is active in the CDISC community, contributing to the development of ADaM and the Statistical Domain Analysis Model, and he serves as a CDISC ADaM trainer. He received an MBA from James Madison University, and a BS in Statistics from Virginia Polytechnic Institute and State University.

Learn more about these authors by visiting their author pages, where you can download free chapters, access example code and data, read the latest reviews, get updates, and more:

- http://support.sas.com/hollandc
- http://support.sas.com/shostak
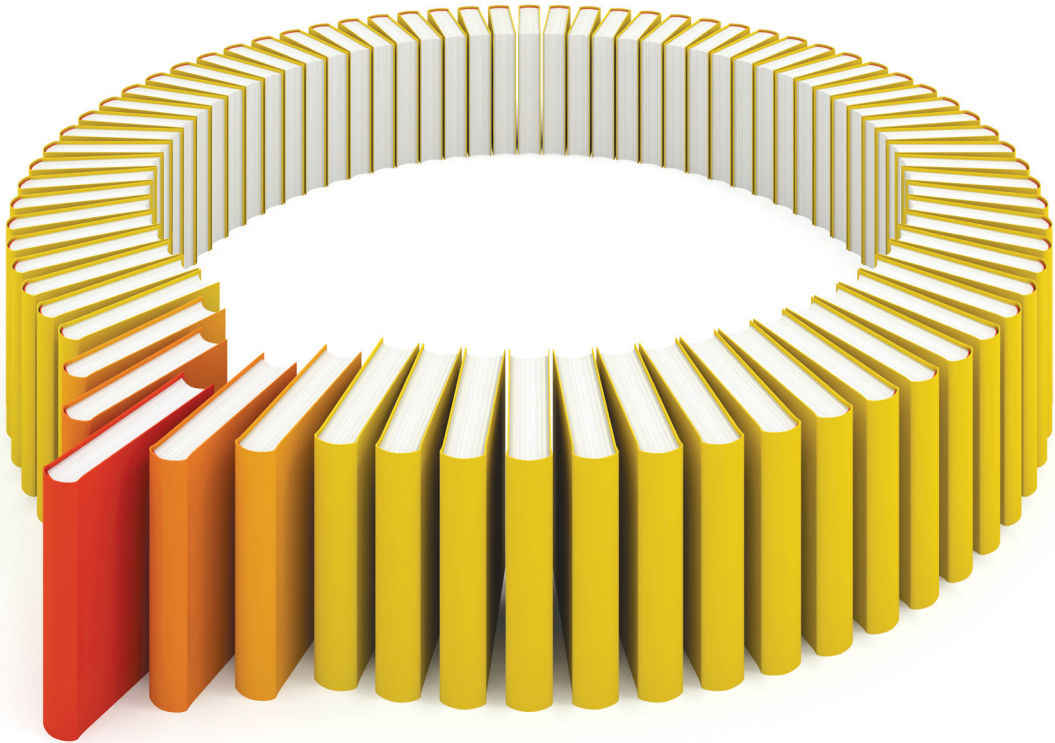
Gain Greater Insight into Your SAS® Software with SAS Books.

Discover all that you need on your journey to knowledge and empowerment.

support.sas.com/bookstore
for additional books and resources.

§sas
THE POWER TO KNOW®