PharmaSUG 2016 – Paper IB10

Moving from Data Collection to Data Visualization and Analytics: Leveraging CDISC SDTM Standards to Support Data Marts

Steve Kirby, JD, MS, Chiltern, King of Prussia, Pennsylvania
Terek Peterson, MBA, Chiltern, King of Prussia, Pennsylvania

ABSTRACT

Data from clinical trials supports a wide range of clinical, safety, regulatory, and analytic groups who all share the same basic need: to efficiently access, analyze and review the data. When clinical data from multiple studies are combined into a "data mart" and linked to visualization and analytical tools, data consumers are able to efficiently find the information they need to make informed decisions.

The raw data as collected in individual studies will vary (at a minimum) based on the specific collection system and forms used. Due to that variability, a foundational step in creating a data mart is to ensure that the data from across studies has a consistent, standard format.

We will share our experience leveraging CDISC SDTM standards to support data marts containing data from many studies across several therapeutic areas. Practical considerations related to ensuring 1) that the SDTM implementation is consistent across studies, 2) that the data made available will support all consumer needs, and 3) that the data will be made available as needed by the consumers will be discussed. Thoughts on how the industry shift towards integrating CDASH standards into collection forms will benefit the future state of visualizations and analytics based on data marts will be shared.

INTRODUCTION

"Standardizing study data makes the data more useful. Data that are standardized are easier to understand, analyze, review, and synthesize in an integrated manner in a single study or multiple studies, thereby enabling more effective . . . decisions. Standardized data also facilitate data exchange and sharing (e.g., between a contract research organization and a sponsor)." [Guidance for Industry Providing Regulatory Submissions in Electronic Format — Standardized Study Data, Draft, February 2012].

Clinical research is driven by data. Data from clinical trials supports a wide range of clinical, safety, regulatory, and analytic groups who all share the same basic need: to efficiently access, analyze and review the data. That common need has increasingly led to efforts to leverage data from clinical studies using visualization and analytical tools that supply simple and powerful access to pooled study data.

DATA MART BASICS AND BENEFITS

For the purpose of this paper, a "data mart" is a database that integrates standardized data from clinical trials across studies and therapeutic areas for a single company. When clinical data from multiple studies are combined and linked to visualization and analytical tools, data consumers are able to efficiently find the information they need to make informed decisions. With proper planning, study data can be effectively integrated from completed studies and (in real time or near real time) from ongoing studies. Raw data are standardized for individual studies, with the standardization done consistently across all studies. After standardization, the data are pooled with data from other studies in the data mart. A simple representation of the concept is in Figure 1 on the next page.

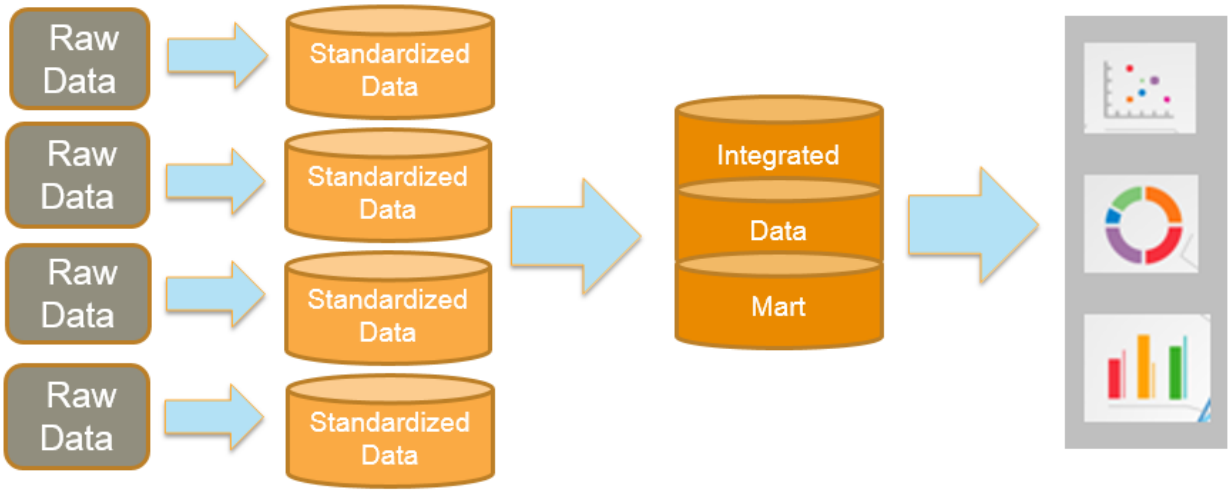

Figure 1. Simple Representation of Data Mart Creation

With standardized study data, or put another way, stable inputs, analytical and graphical tools can be built, used and reused. With those tools in place, consumers can zoom in or out within and across studies, sites and subjects to detect patterns and signals that are hidden from view in a traditional data environment. Such easy, powerful access can simplify study monitoring, streamline review tasks, guide scientific decisions and support a variety of business needs.

DATA STREAM PLANNING

The raw data as collected in individual studies will vary (at a minimum) based on the specific collection system and forms used. Due to that variability, a foundational step in creating a data mart is to ensure that the data from across studies has a consistent, standard format.

The initial decision is whether to use CDISC SDTM as the basis for standardization for the data mart or to follow a custom standardization strategy. If SDTM is not used as the basis for data mart standardization, there will need to be two data streams (as shown below): one for the data mart, and one for submission as shown in Figure 2 below.

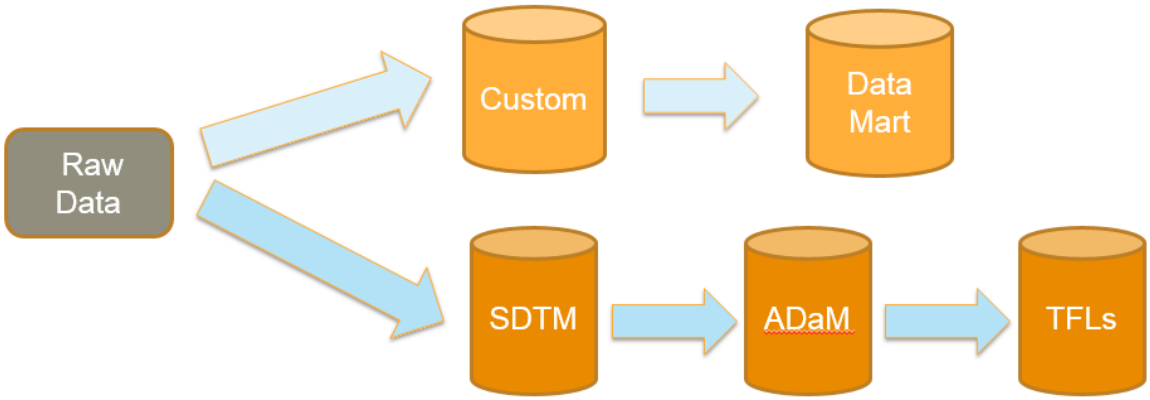

Figure 2. Custom and SDTM Standardization, Two Data Streams

Using a custom approach (not based on SDTM) for the data mart may allow more flexibility to meet the needs of data mart consumers; and having separate streams for the data mart and for submission, analysis and reporting can avoid planning and coordination issues associated with having the same mapping serve two different consumer groups.

Using CDISC SDTM as the basis of standardization for both the data mart and (as needed/required) for analysis, reporting and submission, only one data stream is needed. A simple representation of that is in Figure 3 on the next page.

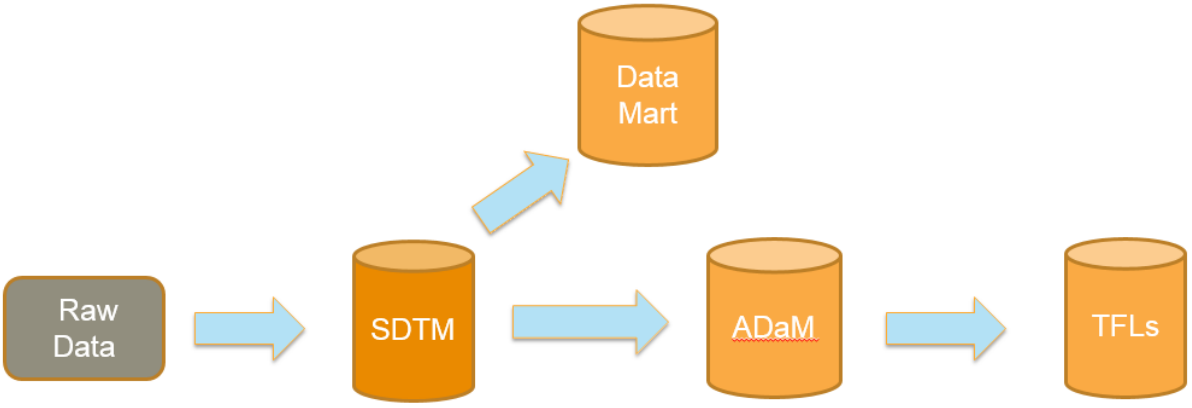

Figure 3. Standardization Based Only on SDTM, One Data Stream

With a single data stream based on SDTM feeding both the data mart and submission, analysis and reporting, data standardization is only done once, the data used by each consumer group is the same, the standardization tends to be more complete and comprehensive as the common reference is a global, comprehensive standard, and the cost tends to be lower (doing the work once is cheaper than doing it twice).

The single data stream approach is the most common (and increasingly popular) approach. When the data needed for the data mart and for analysis and submission are identical the data stream is clean, simple and most easily managed. We will assume a single data stream based on SDTM for the rest of this paper, and will share a few implementation details where indicated in response to common data consumer needs that make it so the data needed for the data mart contains more content than will be used for analysis and submission.

SCOPE OF INPUTS FOR DATA MART BASICS AND BENEFITS

What data consumers need from a data mart and for analysis and submission often differ, so it is important to fully understand the needs of all the consumer groups before determining how best to meet them. Analysis and submission needs are more consistent; the SDTM data needs to fully reflect the conduct of the study. If data mart consumers plan to use the data for data review, or to otherwise assess the quality of the collected data, more information likely will be needed for the data mart than for submission. For example, eCRF system variables may be needed to establish a clear link back to the collection forms but would not be included in the submission data; or differently formatted versions of variables maybe be needed to facilitate use.

When the data mart needs to contain more information than will be included for submission (when the data mart needs to contain SDTM “plus” other content as needed for internal data review and evaluation, content in addition to that in the domains and supplemental domains) those needs must to be built into the standardization process. That additional “plus” information can be added after the SDTM data for submission is generated or the “SDTM plus” data can be generated as an initial step. Both approaches ensure that the SDTM content used for the data mart and for submission and analysis are consistent, and which is chosen tends to be driven by what fits best based on local needs and preferences.

When SDTM plus as needed for the data mart is is created on the backend following creation of SDTM as needed for analysis and submission, the SDTM data mapping processes can be done and maintained just as they would be if serving the data mart was not part of the mix. This tends to make it so SDTM data are available to support analysis needs as quickly as possible and supports a clean, simple path from collection to analysis and submission as shown in figure 4 on the next page.

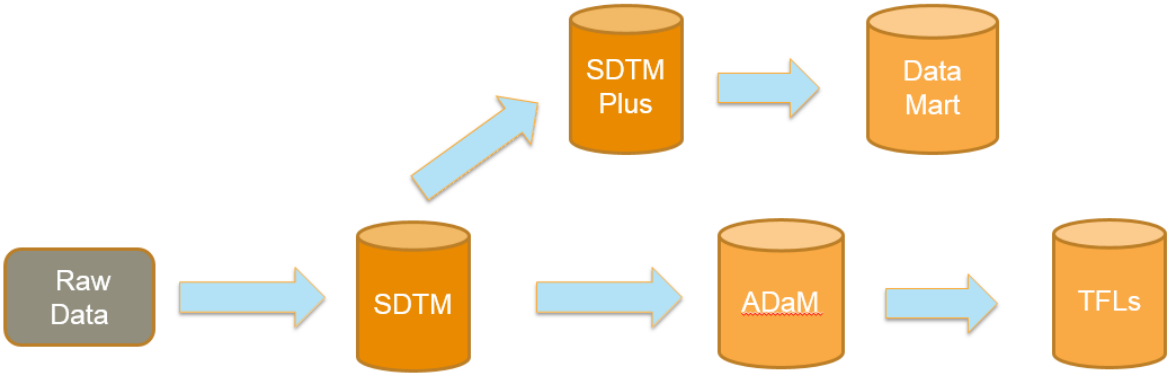

Figure 4. SDTM Standardization, SDTM Plus Created after SDTM

But as part of maintaining a clean, simple path to analysis and submission the path to the data mart becomes more complex. This complexity largely is due to the fact that with this approach the SDTM plus input data is SDTM as needed for analysis and submission as opposed to the raw data as collected. Finding a simple, robust way to merge additional raw data content with SDTM can be challenging and careful planning is needed to avoid awkward complications. In some circumstances, content available in the raw data would be mapped for consistency with SDTM and then the mapped back again to have the data as needed for the data mart, in effect duplicating the risk and burden of mapping. Additionally, some of the content needed for the data mart will live in supplemental domains. That supplemental content would only be included in SDTM for submission and analysis when present which can make supporting null variables and providing a consistent input structure for use in the data mart challenging.

When SDTM plus as needed for the data mart is created as an initial step, the path to the data mart is streamlined and versatile. To generate SDTM as needed for analysis and submission, the SDTM plus data would be subset and content placed in supplemental domains as needed. This approach is shown in Figure 5 below.

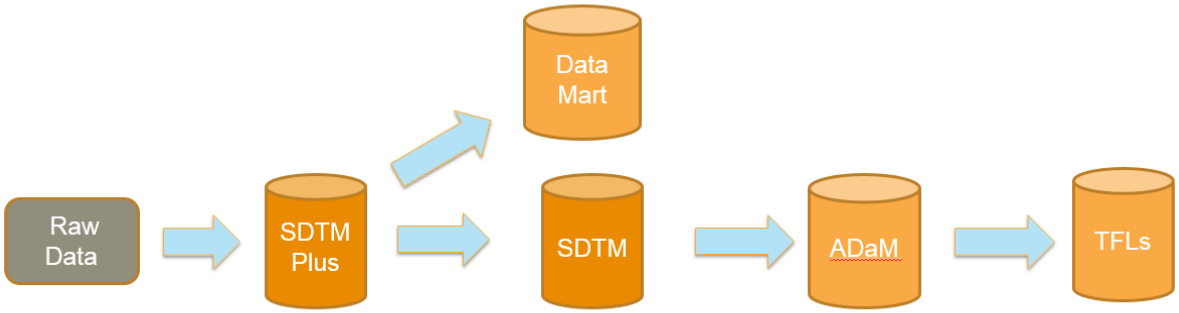

Figure 5. SDTM Standardization, SDTM Plus Created before SDTM

With this approach the needs of data mart consumers can be most easily met as the data mart inputs are one mapping step away from the data as collected. Information that will eventually live in supplemental domains can be retained even if all values are null. But with that prioritization comes an additional burden as delays in finalization of SDTM plus can affect analysis deliverables, and SDTM plus must contain a complete, compliant set of SDTM content.

SCOPE OF INPUTS

Many different data consumer groups are likely to leverage a data mart when it is up and running. Understanding what they need and want to evaluate using standardized, pooled data is an important planning step. And the best way to understand their needs is to ask. Reach out the data consumers early and find out whether they need access to both ongoing and completed studies; see if they expect to look at data quality, safety, demographics, compliance, all of the above or all that and more. And while the potential uses of standardized pooled data are almost limitless, resources tend not to be, and understanding which needs are most critical can help establish priorities.

Once the planned uses are established, making sure that everyone is on the same page for exactly what content needs to be delivered to support those uses is time well spent. Different groups tend to categorize data differently and translation issues are a risk that needs to be managed. For example, if the stated goal is to evaluate safety, ask specifically what substantive areas are considered to be related to evaluation of safety across studies and how any study specific will additions be documented. Then, once the substantive content is clearly delineated, a clear, shared understanding of where and how it will live in SDTM is typically needed.

DATA DELIVERY AND ACCESS

Another key part of planning to support a data mart is to ensure that content will be made available to consumers as needed. Often how soon and how often the data need to be delivered is driven by how consumers plan to use it. As an example, if a key use is to assist with data cleaning or to evaluate compliance in ongoing studies, study data will likely need to be available in the data mart soon after the start of study conduct and will need to be refreshed frequently (perhaps nightly). A transfer schedule that starts soon after conduct will tend to be more costly, so it is reasonable to make sure that the cost of expediting the schedule is covered by the benefits.

Refreshing standardized data nightly is a different challenge than monthly; having mapping code (or transformations) in place very soon after the start of study conduct is a different challenge than waiting until all data from several subjects are available. Understanding the data delivery schedule early makes it so planning to can comfortably be in place to meet that schedule. Without that planning, timelines will likely to be missed. For example, delivery of SDTM very soon after the start of conduct (even when the collection format is not stable across studies) is manageable with a plan in place to leverage dummy data; without that planning, it likely will not happen.

As well, there can be many practical challenges associated with where the data will live, how it will get to its home and who can visit it. The data must be housed in a secure environment with controlled access. The standardized data may be provided by multiple vendors, and applications used to leverage the data mart need to be developed, updated and maintained. A clear, scalable plan to move, house and access the data is needed for the benefits of a data mart to be realized.

PLANNING TO CONSISTENTLY IMPLEMENT SDTM

A data mart will not work unless that data that feeds it is consistently standardized. The raw data as collected in individual studies will vary (at a minimum) based on the specific collection system and forms used. So a critical part of planning to support a data mart is to have a plan in place to consistently implement SDTM. Reasonable minds can differ at the margins about how to map content into SDTM. Collection standards are often not written consistently across studies and therapeutic areas; and they typically do not match up with CDISC standards as well as they should. One way to effectively manage those standardization challenges is to establish (and maintain) local mapping rules. An effective way to do that is by creating and supporting a local SDTM interpretation guide.

The key part of a local SDTM IG is to establish guidelines for how mapping is to be completed. Figure 6 on the next page shows a simple example of that content. These guides can be used by employees and can be provided to multiple vendors to help ensure the SDTM implementation is consistent regardless of who is providing the resources. Additional detail can be added to the guide where useful and manageable in your local environment. From there it is often useful to add additional context by showing exactly how mapping would be done pursuant to the local IG using a representative study.

DOMAIN	ORDER	KEEP_VAR	VARIABLE	LABEL	GUIDANCE
AE	1	Y	STUDYID	Study Identifier	Must match SDTM DM.studyid
AE	2	Y	DOMAIN	Domain Abbreviation	Set to AE
AE	3	Y	USUBJID	Unique Subject Identifier	Must match SDTM DM.usubjid
AE	4	Y	AESEQ	Sequence Number	Based on "keys" variables in the domains tab; should define unique records within a subject
AE	7	Y	AESPID	Sponsor-Defined Identifier	Populate with the AE number from the CRF or related eDC system variable.
AE	8	Y	AETERM	Reported Term for the Adverse Event	Verbatim term from CRF. Note: Only events specifically designated in the protocol or similar as Adverse Events go in this domain.
AE	9	N	AEMODIFY	Modified Reported Term	Populate only if needed/used for coding (for example, can be needed when verbatim AE is a combination of terms or AE terms are adjusted to get an exact match with MedDRA).
AE	10	Y	AELLT	Lowest Level Term	Dictionary coding content. (This may be integrated into a single dataset or may be available separately).

Figure 6. Local SDTM Interpretation Guide (IG)

As a first approximation, all raw content that can be standardized should be. Local rules tend to be most critical where CDISC does not specify all (or nearly all) the details. For example, without local rules: One ARMCD may be associated with multiple drugs; one QNAM may have many labels (and meanings); unscheduled visits may be grouped together in some studies, slotted between visits in others.

It is also worth mentioning that standards implementation is a process and not an event. Over time inputs vary, CDISC standards evolve and local standards need to be maintained. The acronym G.A.I.N is a concise way to highlight the components of a good standards implementation process.

Gather the Information: Listen, understand and develop goals; evaluate current processes; develop a plan for the best approach.

Adopt the plan: Leverage in-house and industry knowledge to create Local SDTM IG; draft and review documentation; build and test the standard; schedule regular review meetings

Implement the plan: Confirm buy in across all roles/levels of organization so enforcement and transition of standards is successful; provide training.

Nurture/Maintain the Standards: Periodic review ensures standards meet business requirements; create processes/documents to support change requirements.

SUPPORTING SDTM (AND DATA MARTS) WITH CDASH

Best to just say it out loud: SDTM-friendly protocols and CDASH eCRFs make SDTM standardization safer, faster, cheaper, and, well, just better. There is less manipulation of source information, a clearer path from collection to submission and a reduced need for supplemental domains. And with greater harmony between the raw data and SDTM, the data as-collected becomes much closer to SDTM plus, and the data stream needed to serve the data mart needs of many data consumers becomes even more linear as shown in Figure 7 on the next page.

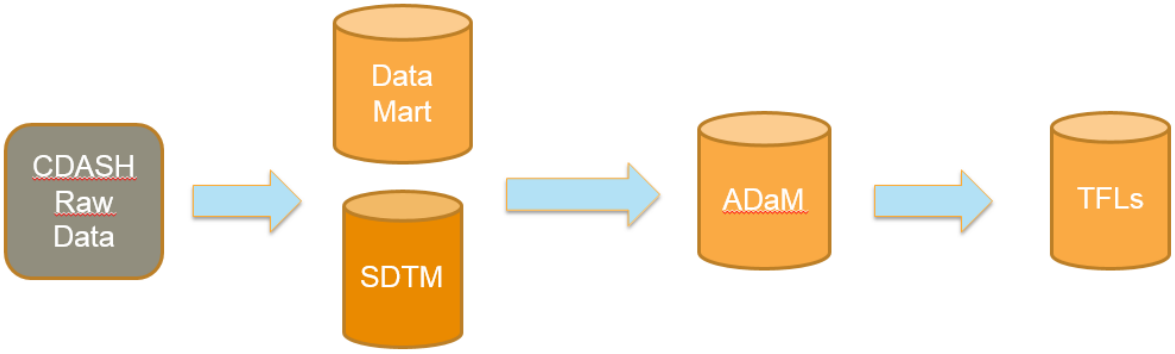

Figure 7. SDTM standardization with CDASH

One of the most effective strategies to support a data mart is to work (with CDASH) to collect what you submit. If CDASH conventions are followed (where applicable) for each variable collected on an eCRF, many variables will directly map into SDTM, and conversion to SDTM can be further simplified by leveraging mapping associations available from CDISC. With support from CDASH collection standards, a local interpretation guide can be generated, supported and used with less effort.

DATA MART VISUALIZATION

Standardized data are more useful. With data from studies standardized and pooled in a data mart, visualization tools can be efficiently created then used and reused. Many different types of data consumers can view and review the data using accessible user interfaces. Figure 8 below shows a couple simple, useful graphs that were created based on standardized data.

Figure 8. Data Mart Visualization: Example Graphs

With data standardized and pooled, and with visualization tools in place, consumers can zoom in or out within and across studies, sites and subjects to detect patterns and signals that are hidden from view in a traditional data environment. Such easy, powerful access can simplify study monitoring, streamline review tasks, guide scientific decisions and support a variety of business needs.

CONCLUSION

Standardization makes data more useful, and a data mart linked to powerful visualization and analytical tools is a great way to help all the data consumers in your company see just how useful standardized data can be.

Basing data standardization for a data mart on SDTM will often lead to efficiencies as SDTM data are already needed for analysis and submission; and creating (and maintaining) a local SDTM IG is one practical way to ensure that the SDTM data are consistently generated across studies, therapeutic areas and vendors. Planning for SDTM before you collect the data with SDTM-friendly protocols and CDASH eCRFs will further reduce the challenges (and cost) of data standardization.

Reach out to the data consumers at your company. Discuss what they need. Then show them how a data mart can help meet those needs.

AKNOWLEDGMENTS

The authors would like to thank Chiltern International Limited for supporting CDISC standards implementation (and this paper). We would also like to thank the CDISC community. Without the CDISC organization and the time and effort of its many, many volunteers, the data standards we use to streamline the drug discovery process would not exist.

REFERENCE

FDA, “Guidance for Industry Providing Regulatory Submissions in Electronic Format — Standardized Study Data [DRAFT].” FDA. FDA.gov. February, 2012. Available at: http://www.fda.gov/downloads/BiologicsBloodVaccines/GuidanceComplianceRegulatoryInformation/Guidances/General/UCM292169.pdf

CONTACT INFORMATION

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

Name: Steven Kirby, JD, MS
Enterprise: Chiltern
Work Phone: 484-319-9011
E-mail: Steven.Kirby@Chiltern.com
Web: www.chiltern.com

Name: Terek Peterson, MBA
Enterprise: Chiltern
Work Phone: (484) 560-8960
E-mail: Terek.Peterson@Chiltern.com
Web: www.chiltern.com