Statistical Computing Environments and the Practice of Statistics in the Biopharmaceutical Industry

Alan Hopkins, PhD
Senior Director, Biometrics, Theravance, Inc., South San Francisco, California, and PharmaStat LLC, Newark, California

Susan Duke, MS, MS
Associate Director, Biostatistics Development Partners, Drug Development Sciences, GlaxoSmithKline, Research Triangle Park, North Carolina

Sue Dubman, MA
Senior Director, Standards and Architecture, Biomedical Data Sciences and Informatics, Genzyme Corporation, Cambridge, Massachusetts

Key Words
Good statistical practice; Statistical operations; Data standards; Metadata; IT architecture; Statistical programming

Correspondence Address
Alan Hopkins, Theravance, Inc., 901 Gateway Blvd., South San Francisco, CA 94080 (email: ahopkins@theravance.com).

The Drug Information Association designates this educational activity for a maximum of 1 AMA PRA Category 1 Credit(s)TM. Physicians should only claim credit commensurate with the extent of their participation in the activity.

If you would like to receive a statement of credit, you must review the article, and complete the posttest and evaluation included on the DIA website. Participants must receive a passing score of 80% or better on the posttest in order to receive a statement of credit. To access the posttest and evaluation, please visit the DIA website at www.diahome.org, select Educational Offerings, and then select Continuing Education from the drop-down menu, and the My Transcript link. This will take you to the My Transcript page where you will be prompted to sign in using your DIA username and password. Once signed in, you may select the article "Statistical Computing Environments and the Practice of Statistics in the Biopharmaceutical Industry."

You will be prompted to complete the posttest and evaluation. Upon successful completion of the posttest, you will be able to download your statement of credit. If you are not a DIA customer, please contact the DIA office at mytranscript@diahome.org for a registration form. There is no fee to receive your statement of credit.

A structured statistical computing environment (SCE) enhances rigor in operational implementation of statistical analyses of clinical studies through process transparency, allowing reproducibility of results by independent reviewers. Desirable features and associated benefits of an SCE system are described. Minimum SCE requirements discussed in detail consist of a structured programming environment, an operational analysis data repository, and a metadata-driven architecture containing information about data and status of various processes. The metadata provide a foundation for connecting multiple processes and systems, thereby allowing the creation of tools that largely automate the analysis process. Standards drive productivity enhancement for creating statistical deliverables based on metadata obtained from the development plan, protocols, and analysis plans.

Not all the features discussed are available today in commercial systems. In the future, nearly all information about clinical trial analytics can be driven by a standards-based, metadata-driven architecture. To accomplish this goal, metadata need to be available about all the processes used to collect, transform, and analyze the patient data. Further standards development will be necessary to fully describe the entire statistical analysis process.

Learning Objectives

Upon completion of this article, participants should be able to do the following:

- *Describe the elements of good statistical practice that contribute to establishing the credibility of clinical trial results.*
- *Describe the fundamental concept of statistical computing environment (SCE); the SCE as a programming environment; and as a clinical data platform and repository driven by a metadata architecture.*

Target Audience

This article is informative for medical doctors working in the pharmaceutical industry; biostatisticians, statistical programmers, clinical data managers, and IT professionals.

INTRODUCTION

The statistician's role is an important one in clinical drug development because it includes responsibilities in design of the clinical program and studies, data analysis, and interpretation of results. These intellectual activities often take a backseat to the significant time spent addressing operational logistics, thereby limiting the time and attention spent adding value to the intellectual property potential of a company's drug assets.

The practice of statistics in the biopharmaceutical industry has made few operational

Release Date: January 1, 2010
Expiration Date: January 31, 2011
Estimated time to complete the activity: 1 hour

Disclosure
It is Drug Information Association policy that all faculty participating in continuing education activities must disclose to the program audience (1) any real or apparent conflict(s) of interest related to the content of their presentation and (2) discussion of unlabeled or unapproved uses of drugs or medical devices. Specific areas to be considered in this disclosure include the following: grants/research support, consultancy relationships, speakers' bureau participation, significant equity (stock) positions, and sources of honoraria.

Alan Hopkins has disclosed that he has no financial relationships to disclose.
Sue Dubman has disclosed that she has no financial relationships to disclose.
Susan Duke has disclosed that she has no financial relationships to disclose.

Submitted for publication: December 22, 2008
Accepted for publication: April 23, 2009

breakthroughs in the last 20 years compared with other drug development disciplines. Effective integrated and scalable productivity systems are largely missing from statistical operations. There is little interprocess communication between analysis tasks and much work is done manually, which takes time away from the high-value work of science. At the same time, there is an increased need for transparency through traceability of the statistical analysis implementation. Traceability refers to the completeness of the information about every step in a process chain. The Wikipedia definition: Traceability is the ability to chronologically interrelate the uniquely identifiable entities in a way that is verifiable.

Recently, structured statistical computing environments have begun to emerge, driven by regulatory guidance documents, standards, and other factors. These systems typically consist of a structured programming environment, an operational analysis data repository, and a layer of metadata containing information about data and status of various processes. The metadata can provide a foundation for connecting multiple processes and systems, thereby allowing the creation of tools that can help automate the analysis process. A statistical computing environment, as we describe, can contribute to credibility of results through process transparency, enabling reproducibility of statistical analyses by third parties.

Attempts to make clinical studies fully electronic have been ongoing since the 1990s, and the industry now has vendor options for almost every step or component of the clinical study. This clinical research space is still being defined and refined, with some areas more mature than others. We believe this is the right time to describe what, in our judgment, statisticians want and need so that colleagues in our information technology departments and at vendor companies can better understand the problem and solution space for this critical part of drug development.

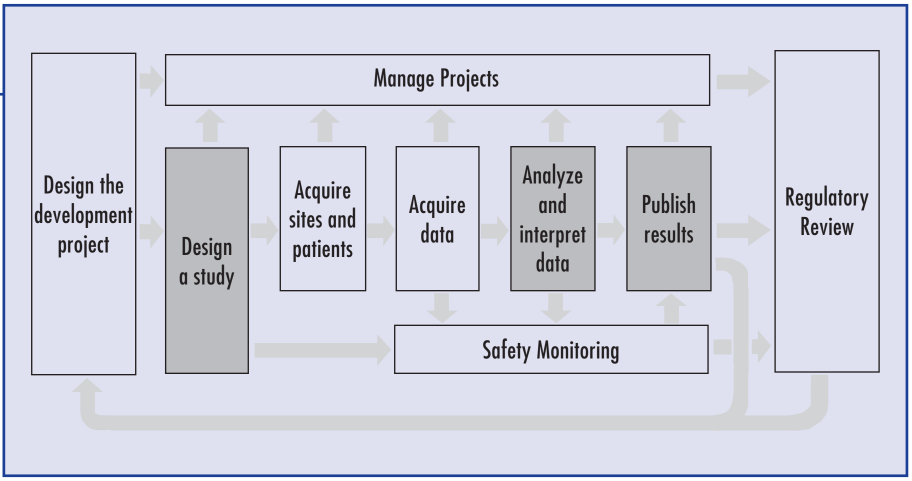
This article describes elements of good statistical practice that contribute to rigorously establishing the credibility of clinical trial results. After describing the fundamental development process and the concept of a statistical computing environment (SCE) in this section, we discuss the SCE as a programming environment. Following that, we discuss how standards can drive productivity enhancement for creating statistical deliverables based on metadata obtained from the development plan, protocols, and analysis plans. Last we discuss the SCE as a clinical data platform and repository driven by a metadata architecture.

DRUG DEVELOPMENT IS DATA INTENSIVE

Clinical drug development is fundamentally based upon data from clinical studies that support the expected or desired drug package label.

FIGURE 1

Generic clinical drug development process.



A good drug development plan identifies a series of clinical trials that collectively produce data necessary to support the safety and efficacy claims for an ideal drug label. A generic representation of the clinical development process is depicted graphically in Figure 1. Each box in Figure 1 either relies on data from prior processes or generates data that will be used later in the process. This requires integration of science, information technology, and statistical practice. Managing the data pathway in clinical drug development should be a core competency among all the participants in the process. This article focuses on the gray boxes, where statisticians spend much of their time.

The protocol and statistical plans for analysis and interpretation of results are important at the study design stage. Information technologies are critical for managing data acquisition, transformation, documentation, storage, and analysis. Information describing the key analysis objects—the metadata—also adds value and needs to be managed. Biostatistics play an important role in these activities from a scientific point of view and operationally—what we call statistical practice. We focus on the practical aspects of analysis and publishing of statistical results in the remainder of the article.

GOOD STATISTICAL PRACTICE

Programming activities for analysis of clinical data are inextricably linked to good statistical practice. ICH E9 (1) discusses statistical principles in drug development and provides a basis for good statistical practice from a regulatory perspective. While the majority of the E9 document stresses good statistical science (minimizing bias and maximizing precision of estimates of treatment effect), documented statistical operations help ensure validity and integrity of prespecified analyses, lending credibility to the results.

In addition to ICH E9, there are multiple other catalysts for a new approach to statistical practice that emphasizes documented reproducible research: Clinical Data Interchange Standards Consortium (CDISC) data standards (2), the CDISC analysis data model (ADaM) guidance (3), HL7 data standards (4), the harmonized CDISC-HL7 information model (BRIDG) (5), electronic records regulations (6), FDA guidance for computerized systems (7), and electronic Common Technical Document (eCTD) data submissions (8). In particular, the ADaM guidance provides metadata standards for data and analyses. This enables statistical reviewers to understand, replicate, explore, confirm, and reuse the data and analyses. The goal is clear: complete and unambiguous communication of decisions, analysis, and results across the clinical data life cycle.

We interpret good statistical practice as a transparent, reproducible, efficient, and validated approach to designing studies and to acquiring, analyzing, and interpreting clinical data. Reproducible research depends upon process transparency and also provides auditability of the statistical analysis. Analysis transparency requires that a navigable electronic process chain exists from defining the objective of the analysis to creating the results, as depicted in Figure 2 (9).

The statistician makes a number of judgments at each stage of this process. These decisions are recorded in the analysis plan, derivation definitions, analysis code, and other forms of metadata. These records of decisions provide transparency that allows the regulatory community to review statistical analyses and constitute good business practice, which may be of partic-

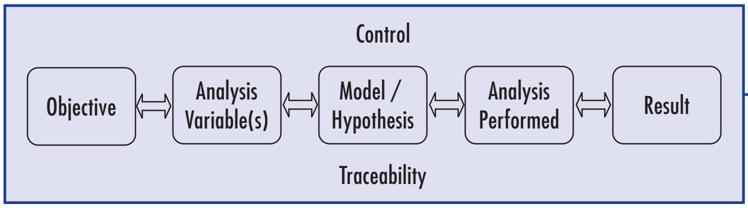


FIGURE 2

Process chain stages for defining statistical results.

ular benefit when further questions about the analysis arise months or years later. The FDA electronic submission guidance defines what FDA expects to receive: multiple types of data files, documentation, and programs—the major components of an analysis environment.

SHORTCOMINGS OF CURRENT BIOSTATISTICAL PRACTICES IN INDUSTRY

Analysis of clinical research data may involve thousands of documents, data files, and programming modules that are logically interrelated in complex ways. Most of these entities are modified by manual and programmatic processes in the course of a project, as data are gathered and cleaned and as analysis requirements are defined and executed. Statisticians spend significant time managing these moving parts: creating files, tracking their status, determining the effect of changes on other parts of the system, and determining that each component meets its requirements and has been verified to be correct. The complexity of such a system makes quality assurance of the whole a huge task. Lack of productivity tools to support basic business processes and associated statistical deliverables creates a broken link in the chain of custody of the data and documentation of statistical analyses, making it difficult for a reviewer to reproduce results.

Good statistical practice dictates that a statistical analysis plan (SAP) should be prepared for clinical studies (ICH E6, ref. 10; E9, ref. 1). The SAP describes the statistical methods planned in more detail than in the study protocol. Nowhere is the format or specific content of these analysis plans described; a standard template is needed across the industry. To fully document and drive activities downstream, the analysis plan needs to have a data layer that provides analysis documentation in a machine-readable format in order to implement the process chain in Figure 2. Without metadata specifications to describe analyses in a very granular way, statistical operations most often rely on a patchwork of processes and tools including Excel spreadsheets, special-purpose macros, and operating system scripts. Although existing systems successfully produce deliverables for regulatory submissions, they lack the architecture and tools that are necessary to provide effective management of the whole process.

Another need is a tool for specification of statistical tables that can be read electronically and used to produce the desired display of information. Typically, specifications for data displays take the form of mock tables created by text-based systems and annotated with programming instructions. The process is tedious and the content is not parameterized or readable by other programs. One issue for the industry is that no software exists to easily specify and create statistical tables of moderate or high complexity. Hopkins and Collins (11) showed conceptually one way to solve this problem based on a particular SAS reporting system. These shortcomings can be addressed as an add-on tool within an SCE.

One of the most conspicuous areas where productivity tools are missing is the actual creation of statistical tables. Extensive programming is often required to assemble and print tables and figures. Data from multiple sources need to be transformed into analysis-ready data. Statistical information derived from different statistical analysis methods needs to be combined for presentation. A common way to address these needs is to build a suite of software macros designed to integrate statistical results and produce the required statistical information. Use of macros, however, requires further programming, validation, support, and maintenance. The reporting tools should be validated, versatile, and accessible to statisticians and not just programmers.

WHAT IS A STATISTICAL COMPUTING ENVIRONMENT?

An SCE is a system that provides a foundation for documenting rigor in the analysis and reporting of clinical trial results while increasing productivity.

Rigor requires transparency, reproducibility, and adequate documentation. The environment should provide role-based security and audit trails for transparency. Reproducibility implies that reviewers can re-create results by being able to execute the same computation or create a computation execution based on a complete and unambiguous specification of the computation. There are two areas of analysis programming that need transparency: the creation of derived variables or observations and analysis or model specifications. Availability of analysis programs that created the computations may also facilitate review.

A productivity system is a comprehensive environment that provides tools to create end products of the analysis process. End products include statistical tables, listings, figures, and documented analysis files. These end products are preceded by in-process SAPs based on the study protocol, case report form, programming specifications, data file documentation, and so on. Tools for creating the end products will depend upon a metadata repository containing all the information needed for communications between tasks. The system should be defined sufficiently so that the process time can be measured and compared with other productivity models and the SCE should be validated for its intended objectives. Only by having defined processes and associated metadata can day-to-day quality and efficiency be easily achieved.

SCE AS A PROGRAMMING ENVIRONMENT: FEATURES AND BENEFITS

As a programming environment, we believe the SCE should be a closed system to allow management of security-related issues, metadata associated with the programming process, the programs and reports, and enforcement of business rules. Such a system can be tailored to both utilize and produce metadata linking the programming requirements from the SAP and mock table shells into ADaM analysis results—metadata.

The programming world has developed many tools for productivity of programmers but these largely have not been assimilated within the biopharmaceutical clinical trials domain. There are six fundamental needs of a statistical programming environment:

1. Security: controlled access and permissions for objects and actions
2. Version control: for traceability and process control
3. Dependency management of objects: a fundamental organizational principle for programming projects
4. Metadata management: enables interoperability and interchange
5. Easy-to-use development environment: analysis tools that are versatile, accessible, and validated
6. Configurable: ability to define and enforce business process rules

The security requirements are established by 21 CFR Part 11, the regulation that directs organizations to have their electronic processes under control; this is just good business practice and supports good statistical practice. The computer operating system could control access and permissions, but it is hard to design a file system with complicated permissions of the sort that might be required when multiple roles exist among different members of the study project team.

Version control of programs, data, and output is needed to control process and provide transparency. It follows that each table and figure has a documented production history with audit trails that capture the reason for changes in program code. If there are multiple releases of results during the study analysis process, what might appear to be discrepant results can be easily tracked based on program audit trails. Version control of the output includes all the input data, production libraries, and program logs. Program status flags, a type of program metadata, can also be used to track the stage of program development and validation status, which is useful for management of large projects.

Dependency management of inputs and outputs is necessary because of the complexity of multilevel steps and the often ad hoc nature of data analysis. For example, a single statistical table might be created using multiple analysis data sets that have been created by other programs from the raw data. There may be multiple files of code that contribute to the creation of the statistical table—macro libraries, initialization files, spreadsheet data, and so on. A final statistical table can only be created (or re-executed) after all its predecessors were created. The computer system should track all dependencies and alert users of stale output deliverables that need to be updated because predecessor dependencies have been updated. Dependency management of statistical output is analogous to managing computer software development through the Unix Make command. With many predecessors, changes to data or programs themselves will often result in many tables needing to be rerun. For this reason, the environment needs to make it easy to rerun entire batches of jobs without user intervention. This requirement for production systems to be batch-oriented distinguishes production systems from program development activities that are usually interactive and iterative.

Metadata management is needed to ensure consistency in use and meaning of content within the SCE and across the clinical data life cycle, from trial design through submission and beyond. Metadata includes clear, unambiguous data element definitions used in the SCE as well as in interfacing systems. In the most general sense, metadata answer who, what, when, where, why, and how about every facet of the process and study data. Some metadata (like data definitions used in case report forms) will likely come from an external master metadata repository. However, the SCE itself is a major producer of metadata needed to manage statistical processes and work flow so must incorporate metadata management capabilities within its own environment.

Ease of use is critical to user acceptance of the new working environment. Early program development and the iterative changes associated with making a working program stable could be done outside the SCE to avoid version control when it may not be most useful. Once programs start producing output shared with a project team, business rules typically kick in and require that the programs reside in the SCE. Then further changes are subject to version control. Simple rules like this enhance user acceptance.

Configurability permits customization based on existing business rules, for example, uploading validated tables and figures into an outside authoring environment automatically once validation has occurred. These and other desirable features from an SCE programming point of view are enumerated in Table 1.

The archive is an important function that allows transportability of the work environment that created the statistical analyses. The archive should be system independent. The archive is needed for review and use by others, such as a data monitoring committee (DMC) and development partners, as well as for established records management requirements. A sponsor could program prespecified DMC reports and export them to an independent statistician who would run the programs by adding the unblinded treatment codes. The SCE needs to be open and extensible so that statisticians need not rely on a single vendor's analysis software. The environment must allow the statistician to do analysis over the network or the web, on a local machine, or through the SCE environment servers.

As with deployment of any new tool, process realignment must be considered. An SCE will not be able to reach its full business value if implemented without some level of established and enforced statistical programming business process. In order for an SCE to be successful there needs to be a harmonization of people, business process, and technology.

SCE: A PRODUCTIVITY ENVIRONMENT FOR STATISTICAL DELIVERABLES

The SCE vision is to improve statistical capability and productivity by making nearly all information about clinical trial analytics meaningfully electronic. We propose use of formal electronic representations of concepts connected to one another instead of unstructured content scattered over multiple documents. To implement the vision operationally, two key components are required: data standards and software tools. Multiple data standards and tools can be connected together through interoperable metadata—information about the connections between the data, its analysis, and tools.

TABLE 1

Desirable Features and Associated Benefits of a Statistical Programming Environment

Feature	Benefit
Process automation	Create repeatable, enforceable, predictable processes
Version control and accountability	History of all documents including statistical analysis source code, data files, logs, and output, and establishes chain of custody
Dependency management	Impact analysis and update process management (further details under SCE Architectural Considerations)
Document repository	Store documents associated with programs, data, and analysis
Metadata management	Ensure consistency in use and meaning of data and analysis; manage all metadata used in processing of information, work flow, and change control
Submission metadata publishing	Create reports such as define.xml data documentation, automated tables of contents, footnotes, and titles
Program status flags	Allow enforcement of business rules concerning validation
Reporting and metrics	Improve visibility about projects and processes
Work flow	Configure work flows for communications between tasks
Information exchange	Allow for the SCE to exchange information with other systems, both internally and with external partners and vendors, and to predictably use the information that has been exchanged
Local, web, and SCE access	Access from multiple environments and locations
Security	User authentication and user authorization to ensure changes are made only by authorized individuals; role and life cycle–based security model with all security activities tracked in the audit trail
Electronic approval management including electronic signatures	Simulates a legally binding signing process and allows individuals in an approval chain to add information and sign their names just as they would on paper without invalidating the previously applied signatures
Audit trails	Trace the origin and detail of all activities
Archive	Export the environment from the repository or manage an internal archive process and artifacts; ensure conformance to established archive rules and standards
Extensibility	Must be able to run multiple statistical packages, eg, SAS, R, S-PLUS, etc and multiple versions of supported packages due to the fact that the clinical program life cycle will often run longer than a particular version of a software product
Scalability	Ability to support small or large organizations

This section presents process implications for statisticians of some of the production and emerging CDISC standards and describes productivity tools that can be enabled by standards.

THE IMPACT OF STANDARDS

Data standards are foundational for creating an environment where tools can be leveraged at different points in the analysis process. Standards for clinical development of drugs have been defined and are maturing under various initiatives of CDISC, the HL7 RCRIM Working Group, and the BRIDG initiative. “The BRIDG Model is a collaborative effort of stakeholders from the Clinical Data Interchange Standards Consortium (CDISC), the HL7 Regulated Clinical Research Information Management Technical Committee (RCRIM TC), the National Cancer Institute (NCI), and the US Food and Drug Administration (FDA) to produce a shared view of the dynamic and static semantics that collectively define a shared domain-of-interest, i.e. the domain of clinical and pre-clinical protocol-driven research and its associated regulatory artifacts” (5). Very important to all of this are the standard structures by which we communicate the physical data: the CDISC Operational Data Model (ODM), HL7 v3 XML Messages based on BRIDG, and the Study Data Tabulation Model (SDTM). Tools will need to become data structure–aware in order to create the downstream deliverables.

Controlled terminology including CDISC standard terminology (12) as well as MedDRA, SNOMED, and other vocabularies provide a shared way of communicating about objects we are manipulating. This aspect of semantic interoperability is part of what will be needed to allow multiple tools to work together in a manner that is understandable and unambiguous. In addition, we will need a common information model (such as BRIDG) that provides understanding of the relationships among information so its meaning can be understood in the context within which it is used as well as a foundation of rigorously defined data types (such as those defined by ISO and HL7) and formal processes and tools for defining interchange structures.

The CDISC-HL7 Protocol Representation standard (PR), being incorporated into BRIDG, is also very important to drive statistical analysis activities downstream. PR data structures such as the Trial Design Model (TDM) describe the study design and the schedule of activities while other structures capture inclusion and exclusion criteria, the core objectives and associated analysis variables models that are to be applied to the collected data. The PR can then serve as input for development of a structured SAP.

Written documentation of the analysis data and the analyses themselves are recommended in the CDISC ADaM guidance (3). This type of documentation goes a long way toward making the process of reporting statistical results transparent. Inside a well-designed statistical environment, the reporting metadata allow streamlining the creation of the ADaM documentation.

METADATA ENABLE WORK FLOW TOOLS

Metadata consist of both global and study-specific data element definition standards, other content information (eg, analysis specifications), and process (work flow) information. Every object that is managed has metadata, so it can be determined how the pieces are hooked together. Instead of islands of information, there are now intrinsic connections through metadata that define key dependencies. When an input data set or specification changes, the system can detect which output objects become stale and need to be refreshed. A metadata manager for tracking input and analysis metadata is a central tool that needs to be an integral part of the SCE. The tools necessary to manage analysis work flow are enumerated in Table 2.

Beyond the metadata manager, trial design information needs to be captured. If one does not have a structured protocol writing tool, the SCE could have a simple study designer to create the SDTM TDM data structures for use by downstream tools such as a table designer. The table designer stores specifications for tables and figures in a computer-readable format like XML (eXtensible Markup Language). The table specifications and SAP can be held in a master metadata repository (as in Figure 3) or inside the SCE.

It is important to note the differences between the SAP and the derivations in the metadata. They are not always and often cannot be the same. The SAP definitions are rarely complete because they do not capture data handling or may make assumptions about data distribution, covariates, and so on. Incompleteness may also be due to the level of detail or granularity of the specification. Data handling changes are basically decisions about processing the data that are not prespecified and can affect the results. It is important to be able to update the derivations without impacting the SAP, and then be able to track those changes for ADaM metadata.

The metadata for the analysis data sets (ie, the initial specifications or definitions) are used for input directly to statistical functions found in statistical software packages. Similarly, the table specifications could be used as input to a report code generator that would create executable computer code to be run in the SCE. In addition to the metadata for the analysis data sets, there are other metadata and data elements for traceability (see ref. 3). So, the universe of analysis data components in the SCE will be larger than the output universe in some respects. There will be variables that do not necessarily match up to output because they are not used in the generation of output. All of these intermediate data-handling specifications

TABLE 2

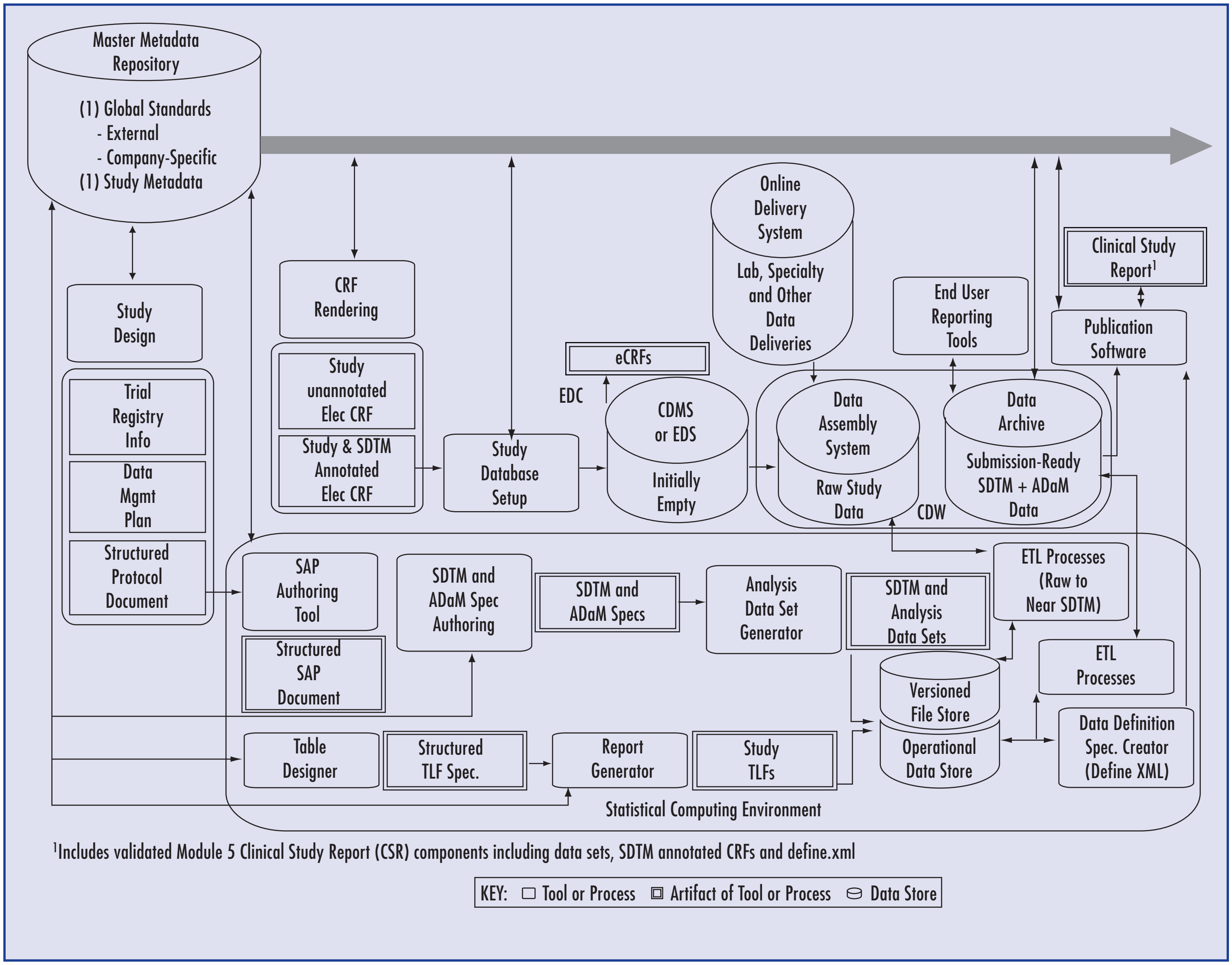
Tools to Automate Statistical Processes

Tool Name	Purpose	Enabling Standards
Metadata manager	The application that utilizes and contributes to information in a master metadata repository as well as collecting and validating metadata from tools or files within the SCE and stores the information in a database for use by other tools	Requires tools with known input requirements and prespecified output. Leverages standards such as HL7v3 RIM, ISO 11179, controlled terminology (eg, SNOMED, MedDRA, etc), and ISO/HL7 data type specifications
Study designer/ protocol builder	Author and/or capture trial design information electronically for use by other tools within the SCE and elsewhere (external to the SCE)	TDM from the CDISC-HL7 study protocol representation
Data architect	Manage and validate study data specifications with data capture standards	Terminology, ODM, clinical data acquisition standards harmonization
Case report form builder	Create a data collection instrument—paper or electronic (external to the SCE)	Clinical data acquisition standards harmonization and ODM XML for data transport now and HL7 messages in the future
Statistical analysis plan builder	Create an electronic representation of the statistical analyses planned	CDISC-structured SAP standard (under development)
Table designer	Create statistical table specifications	ADaM results metadata
Analysis data set code generator	Use table specifications to define data structures required for statistical analysis	ADaM
Report code generator	Create statistical analysis programs to create tables specified in the SAP	Standard software does not exist for creating statistical tables nor is there a standard vocabulary for defining statistical tables
SDTM mapping and transformation manager	The application used to specify and iteratively refine required mapping metadata and transformations of source data to ensure SDTM compliance	SDTM
Data definition specification publisher (Define.xml)	Create documentation for STDM data and statistical analyses	CRT-DDS, ADaM Define.xml, ADaM
Export objects	Transfer reports to an authoring environment	Dependent upon particular authoring tools and document repository

and program characteristics describing what modules were called can be captured as metadata and stored in a study-specific metadata store.

Next, consider the dynamics of metadata: how the metadata repository is populated from the various documents (structured text and data) that use it over the course of a study. How do metadata, data repositories, and structured text documents get populated as a study goes from beginning to end? Figure 3 demonstrates how these files and the metadata they have in common would be populated over a study's life cycle.

If we had these tools, how would our work be different? Authoring processes would evolve to articulate the plan for a particular study in terms of predefined components. Once there is a library of common components with real meaning, the need to create new components should diminish over time as the pieces in this process are relatively standard. Authoring should be more upstream and directly collaborative with other functions by specifying statistical content as fully and as clearly as possible early in the study life cycle (9). This would allow one to leverage the structured content downstream to automate or significantly assist those downstream processes like populating the data capture tools, data set creation, and



[1] Includes validated Module 5 Clinical Study Report (CSR) components including data sets, SDTM annotated CRFs and define.xml

FIGURE 3

SCE workflow in a standards-based, metadata-driven clinical data life cycle environment.

analysis programming. If there needs to be an additional analysis (eg, based on unexpected findings), one needs to be able to record the new plan and record the execution of that plan. That should all be part of the same cycle of defining metadata and executing it, not something in an ad hoc text file with no documentation or context.

SCE ARCHITECTURAL CONSIDERATIONS

SCE AS AN OPERATIONAL DATA STORE

The SCE houses the analysis data, statistical programs, outputs, and other files for the analysis of a specific study. It is the authoritative source for management decisions, submissions, and publications that should be reinforced by policy and standard operating procedures. This collection of files is under management including version control and audit trails, so by definition the SCE is a controlled repository. The SCE becomes a natural place to store all patient data and perform appropriate integration of various data streams. Since the SCE is a repository of files, it is straightforward to import and export information. These files are analysis ready. That is, summary and analytic statistics can be calculated and produce the same results as published results with minimal preprocessing involving simple record selection.

Traditionally, almost all clinical data were maintained in a clinical database management system. Study sponsors would directly enter data from case report forms into the database system, issue queries, integrate external lab information into the database structure, and map certain reported terminology using specialized dictionary standards. Data were exported from the clinical database for analysis. However, in the electronic data capture world we now live in, there are multiple streams of data: electronic case report form data from electronic data capture applications, electronic patient reported outcomes, lab results, randomization information, and so on. These data trails are often asynchronous. It is essential to have a common location for all the data in a study for cleaning purposes as well as subsequent analysis and review. The SCE could be a hub for integrating the clinical data streams.

Unlike a repository, a data warehouse is based upon a formal database structure—for example, a relational table with a schema that defines the structure of the data elements. Information is retrievable through query tools within the database software. We do not see the data warehouse as an integral part of the SCE data hub concept. Integration of study data into a data warehouse archive consisting of many studies with multiple indications or multiple drugs is a value-added activity, but that happens later, not at analysis time. The repository concept is preferable to a data warehouse for operational work because the analysis scenario is too complex and varied to be handled within the structure of a data warehouse.

Finally, the linchpin for efficient and effective usage of data repositories is an associated metadata database. Metadata have always been an important component of the analysis environment because hundreds of files are produced (programs, logs, outputs, and miscellaneous other files) that use similar metadata (eg, protocol number, treatment descriptions, and study visit schedule). The analysis of clinical studies arguably uses more metadata than any other part of the clinical trials process because it is where the thousands of pieces of information from a clinical study are transformed into interpretable results. If an organization does not have an existing master metadata repository, the SCE could be the appropriate place to start to build out the metadata database. Maintaining data in a repository without its companion metadata is a low-value endeavor.

INTEROPERABILITY AND THE METADATA HUB ARCHITECTURE

Much discussion has focused on the pivotal nature of the metadata. Figure 4 depicts how the tools necessary to automate the statistical analysis process are dependent upon a central metadata repository. The metadata manager tool collects metadata and validates it against the standards. With metadata management in place, ADaM variable and value-level analysis metadata can be captured from an SAP. Supplemented metadata that may exist outside the SAP can be integrated and then the analysis and data definition tables can be created for regulatory submission.

Figure 4 is an example of hub-and-spoke architecture. All the tools read and write metadata into the repository. Soloff and Boisvert (13) also recommend this type of architecture because it can:

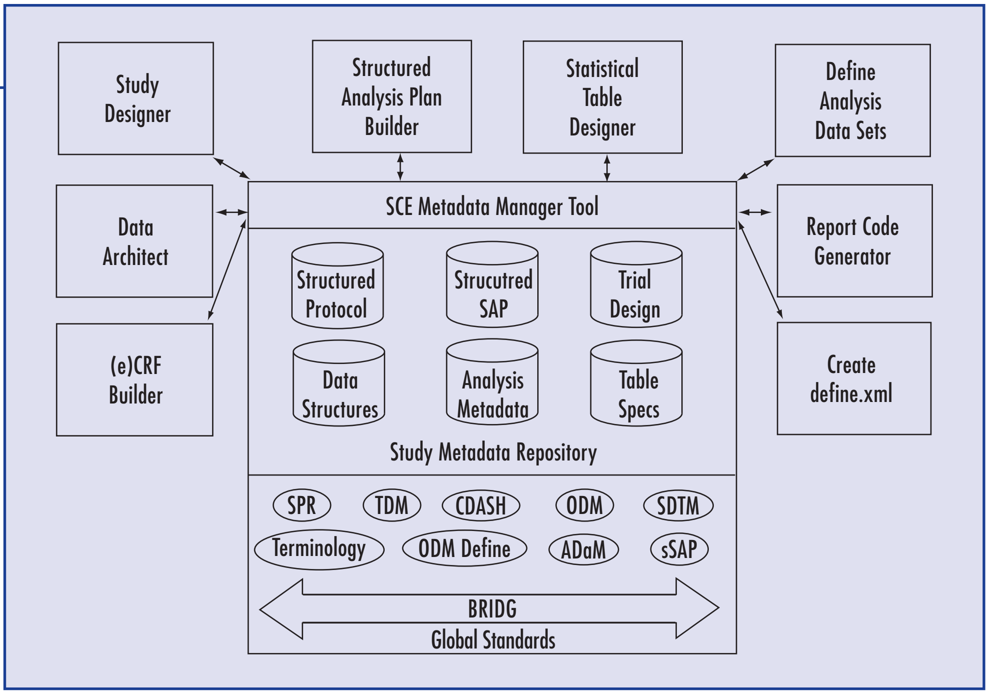
- Govern the process
- Consolidate systems
- Automatically capture and utilize metadata
- Allow for queries to run against this metadata
- Allow for extensible add-ons for future solutions

From the SCE point of view there are really only a couple of inputs that are necessary for our process from the outside: the clinical study data and the protocol design. If both of these inputs are based upon CDISC standards, it will be straightforward to incorporate these data directly into the SCE.

We believe the SCE is the right place to start the architectural transition toward a hub-and-spoke architecture because most metadata relevant to clinical software applications are also needed in the SCE. Once defined and captured within the SCE, the standard interface connectors can also be applied to the other systems in

FIGURE 4

An architecture to support statistical work flow.



the clinical trial life cycle. One way of integrating enterprise systems is by communicating metadata and content from one application to another through asynchronous messages (14). Under this model, one application can poll other applications for data through an established communications protocol. The RCRIM technical group in HL7 is creating transport messages for (a) CDISC study design content, (b) study participation content, (c) CDISC subject data content, and (d) individual case safety reports. Ultimately we expect that HL7 messages will be the mode of communication of information (content) across applications, with CDISC providing the content standards and terminology needed for semantic interoperability.

DISCUSSION

Clinical trials data are the most valuable intellectual property of any biopharmaceutical product development organization. Good clinical practice (10) states that "quality control should be applied to each stage of data handling to ensure that all data are reliable and have been processed correctly." Quality principles stress the incorporation of quality up front in any process. The SCE is proposed as an overall framework for structuring and managing how the raw data are transformed into results through statistical analysis. Therefore the SCE needs to provide full life cycle management for content in the data repository including a documented chain of custody and change control. The SCE can make the data accessible for use by those with responsibility for its collection and analysis. Ideally, connectors will provide seamless import and export with other enterprise systems. In this sense, the SCE becomes a central data hub for information management in the clinical development arena. Tools need to exist within the SCE to update, extract, and view as well as analyze the data.

In the future, it is our hope that nearly all information about clinical trial analytics will be driven by a standards-based, metadata-driven approach. To accomplish this goal, metadata need to be available about all the processes used to collect, transform, and analyze the patient data. New tools will be needed, as will some additional standards, such as an electronic SAP template that parallels the structured protocol representation. So, there is more work to be done to achieve the ultimate process automation, but SCEs are already commercially available that can accomplish several of the tasks identified related to programming and metadata management.

The data standardization efforts of CDISC and the CDISC-HL7 standards harmonization initiatives are encouraging software vendors to create standards-based, interoperable tools and environments that are closer to off-the-shelf solutions that can be used by sponsors, contract research organizations, and regulators. The cost of biopharmaceutical product development includes all the software tools made internally at sponsor companies and CROs and requested from vendors. While each company will have different needs, we all begin (protocol) and end (submission) at the same place—why not take advantage of similar tools? We also ask vendors to consider their roles in the cost of drug development. The blockbuster model in the industry is becoming less common. We need software solutions that are modular and affordable. In the statistics area, open-source software is a strong component of academic statistics programs. We anticipate open-source and lower cost statistics solutions will become more common as vendors who understand our ties to academic colleagues, our need to use multiple vendors' software tools, and the value of open source gain more traction in our industry.

In our discussion we have largely ignored the "Publish Results" box in Figure 1. Seamless integration of statistical results into documents is an important process and one that the industry has not effectively addressed. The literate statistical practice approach (15,16) of embedding computer code into dynamic documents holds promise as a method for completing the chain of documentation and transparency. New tools (perhaps XML authoring software) may also facilitate this approach. Structured documents, from the protocol and statistical analysis plan to the final study report (which all include both data components and text), are necessary components to automate the statistical analysis and reporting process. There are many opportunities to make our analytical and reporting capabilities more efficient as part of the overall process of managing the information life cycle of drug development.

We anticipate substantial change in the statistical environment of the future. Future processes will be based on a prescriptive specification of metadata that will drive process automation downstream. We will be able to observe the process by viewing the resulting metadata that will be connected in a way that elucidates the process that was actually followed. This will make life simpler for both drug developers and reviewers. The standards foundation for this new approach is evolving well. The statistical tools space is lagging but we expect development to be funded by contracts between forward-looking firms and their software partners. We also expect breakthroughs that were not even envisioned when we set out to write this article.

An SCE as we have described makes good sense from both science and business perspectives. Well-defined analyses and well-structured data that facilitate the peer review process in the public health arena are good for everyone.

Acknowledgments—The preparation of this article was inspired by the session Statistical Computing Environments: Establishing the Need, Setting Forth Requirements at the 43rd annual meeting of the Drug Information Association held in Atlanta, Georgia, June 2007, organized by Susan Duke. We recognize the contributions of the session speakers: Greg Anglin, John Brega, Jason Burke, Alan Hopkins, and Steve Wilson. Thanks to Joel Hoffman, Susan Kenny, Wayne Kubick, Becky Kush, Sandor Schoichet, Andy Siegel, and Dana Soloff for their insightful review comments of the draft manuscript. Our thanks go to innumerable other colleagues who have influenced our thinking and who support the concept of a statistical computing environment.

REFERENCES

1. US Food and Drug Administration. Guidance for industry. E9: statistical principles for clinical trials. September 1998. http://www.fda.gov/cder/guidance/ICH_E9-fnl.PDF.

2. Clinical Data Interchanges Standards Consortium. Study data tabulation model. http://www.cdisc.org/standards/index.html.
3. Clinical Data Interchange Standards Consortium. Guide to documents relating to the CDISC analysis data model (ADaM). http://www.cdisc.org/standards/index.html.
4. Health Level Seven (HL7). Version 3.0 Specifications. http://www.hl7.org/about/hl7about.htm#V3.
5. BRIDG (Biomedical Research Integrated Domain Model). http://www.bridgmodel.org.
6. Title 21 of the Code of Federal Regulations. Electronic records; electronic signatures (21 CFR Part 11).
7. US Food and Drug Administration. Guidance for industry: computerized systems used in clinical investigations. May 2007. http://www.fda.gov/cder/guidance/7359fnl.pdf.
8. US Food and Drug Administration. Guidance document: study data specifications, version 1.1. July 2006. http://www.fda.gov/cder/regulatory/ersr/Studydata.pdf.
9. Anglin G. Emerging CDISC standards: process implications for statisticians. Presentation at DIA annual meeting, Atlanta, GA, 2007.
10. International Conference on Harmonization. E6 good clinical practice: consolidated guideline. April 2006. http://www.fda.gov/cder/guidance/959fnl.pdf.
11. Hopkins A, Collins L. Statistical table specifications and automatic code generation using XML. PharmaSUG 2005 proceedings. http://www.pharmastat.com/pdf/TableSpecificationsandAutomaticCodeGeneration.pdf.
12. Haber MW, Kisler BW, Lenzen M, Wright LW. Controlled terminology for clinical research: a collaboration between CDISC and NCI enterprise vocabulary service. *Drug Inf J.* 2007; 41:405–412.
13. Soloff D, Boisvert D. Metadata management in an SCE. Presentation at the DIA CDM/eClinical conference, Washington, DC, 2008.
14. Hohpe G, Woolf B. *Enterprise Integration Patterns: Designing, Building, and Deploying Messaging Systems*. Boston: Addison-Wesley; 2004.
15. Rossini AJ. Literate statistical practice. In Hornik K, Leisch F (Eds.), *Proceedings of the 2nd International Workshop on Distributed Statistical Computing* (DSC 2001). Vienna: Technische Universität Wien; 2001. http://www.ci.tuwien.ac.at/Conferences/DSC.html.
16. Rossini A, Leisch F. Literate statistical practice. University of Washington Working Paper Series, 2003. http://www.bepress.com/uwbiostat/paper194.

The authors report no relevant relationships to disclose.