Guidance for Industry

Providing Regulatory Submissions in Electronic Format — Standardized Study Data

DRAFT GUIDANCE

This guidance document is being distributed for comment purposes only.

Comments and suggestions regarding this draft document should be submitted within 60 days of publication in the *Federal Register* of the notice announcing the availability of the draft guidance. Submit comments to the Division of Dockets Management (HFA-305), Food and Drug Administration, 5630 Fishers Lane, rm. 1061, Rockville, MD 20852. All comments should be identified with the docket number listed in the notice of availability that publishes in the *Federal Register*.

For questions regarding this draft document contact (CDER) Kieu Pham at 301-796-1616, (CBER) Office of Communication, Outreach and Development (OCOD) at 301-827-1800 or 1-800-835-4709, or (CDRH) Terrie Reed at 301-796-6130.

U.S. Department of Health and Human Services
Food and Drug Administration
Center for Drug Evaluation and Research (CDER)
Center for Biologics Evaluation and Research (CBER)
Center for Devices and Radiological Health (CDRH)

February 2012
Electronic Submissions

Guidance for Industry

Providing Regulatory Submissions in Electronic Format — Standardized Study Data

Additional copies are available from:

Office of Communications
Division of Drug Information, WO51, Room 2201
Center for Drug Evaluation and Research
Food and Drug Administration
10903 New Hampshire Ave., Silver Spring, MD 20993-0002
Phone: 301-796-3400; Fax: 301-847-8714
druginfo@fda.hhs.gov
http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/default.htm

or

Office of Communication, Outreach and Development, HFM-40
Center for Biologics Evaluation and Research
Food and Drug Administration
5515 Security Lane, Rockville, MD 20852
http://www.fda.gov/BiologicsBloodVaccines/GuidanceComplianceRegulatoryInformation/Guidances/default.htm
(Tel) 800-835-4709 or 301-827-1800
http://www.fda.gov/cber/guidelines.htm.

or

Office of Communication, Education, and Radiological Programs
Division of Small Manufacturers Assistance, WO66-4613
Center for Devices and Radiological Health
Food and Drug Administration
10903 New Hampshire Avenue, Silver Spring, MD 20993-0002
http://www.fda.gov/Medical Devices/DeviceRegulationandGuidance/GuidanceDocuments.htm
Email: dsmica@cdrh.fda.gov
Fax: 301.847.8149
(Tel) Manufacturers Assistance: 800.638.2041 or 301.796.7100
(Tel) International Staff Phone: 301.827.3993

U.S. Department of Health and Human Services
Food and Drug Administration
Center for Drug Evaluation and Research (CDER)
Center for Biologics Evaluation and Research (CBER)
Center for Devices and Radiological Health (CDRH)

February 2012
Electronic Submissions

TABLE OF CONTENTS

Technical specifications and additional online resources associated with this guidance are provided separately and will be updated periodically. To ensure that you have the most recent versions, check the FDA Study Data Standards Resources Web page at http://www.fda.gov/ForIndustry/DataStandards/StudyDataStandards/default.htm.

Guidance for Industry[1]
Providing Regulatory Submissions in Electronic Format — Standardized Study Data

This draft guidance, when finalized, will represent the Food and Drug Administration's (FDA's) current thinking on this topic. It does not create or confer any rights for or on any person and does not operate to bind FDA or the public. You can use an alternative approach if the approach satisfies the requirements of the applicable statutes and regulations. If you want to discuss an alternative approach, contact the FDA staff responsible for implementing this guidance. If you cannot identify the appropriate FDA staff, call the appropriate number listed on the title page of this guidance.

I. INTRODUCTION

This is one in a series of guidance documents intended to assist sponsors and applicants making regulatory submissions to the Food and Drug Administration (FDA) in electronic format. This guidance establishes FDA's recommendation that sponsors and applicants submit study data in a standardized electronic format.[2] This guidance applies to submissions of clinical and nonclinical study data within investigational new drug applications (INDs), new drug applications (NDAs), abbreviated new drug applications (ANDAs), investigational device exemptions (IDEs), biologics license applications (BLAs), premarketing notifications (510(k)s), and premarketing approval applications (PMAs), including original submissions, amendments, and supplements, to the Center for Drug Evaluation and Research (CDER), the Center for Biologics Evaluation and Research (CBER), and the Center for Devices and Radiological Health (CDRH).

The purpose of this guidance is to promote the use of data standards in the electronic submission of study data to the FDA. We recognize that the use of data standards is complex and this guidance, by itself, is insufficient to provide all the information necessary to implement data standards for study data submissions. Therefore, this guidance refers to a number of technical specifications and other resources associated with the use of data standards that are intended to help submitters determine which data standards to use and how to use them.[3] These technical specifications and resources are referenced online to make them more accessible to the submitters. In some cases, the online resources are maintained by standards development organizations (SDOs) and not by FDA.

---

[1] This guidance has been prepared by the Center for Drug Evaluation and Research (CDER), the Center for Biologics Evaluation and Research (CBER), and the Center for Devices and Radiological Health (CDRH) at the Food and Drug Administration.

[2] For purposes of this guidance, *study data* means both the data collected from experimental subjects (human or animal) in a study and derived data (i.e., the data calculated from the collected data).

[3] For purposes of this guidance, *submitter* means an entity that submits study data to the FDA. Submitters include both sponsors and applicants.

These resources are available on the Study Data Standards Resources Web page at http://www.fda.gov/ForIndustry/DataStandards/StudyDataStandards/default.htm.

FDA's guidance documents, including this guidance, do not establish legally enforceable responsibilities. Instead, guidances describe the Agency's current thinking on a topic and should be viewed only as recommendations, unless specific regulatory or statutory requirements are cited. The use of the word *should* in Agency guidances means that something is suggested or recommended, but not required.

II. BACKGROUND

FDA routinely receives in regulatory submissions the results of scientific studies such as clinical trials and animal studies. The clinical trials are generally human investigations that are intended to study the pharmacokinetics, pharmacodynamics, bioequivalence, safety, and/or effectiveness of an investigational product. For devices evaluated in a 510(k), the studies may compare the products to a *predicate* device. Applicants typically submit study reports, which describe the study protocol, the data collected, the analyses performed, the results of those analyses, and the conclusions of the study. Also accompanying the study reports are the case report forms (CRFs), and the study data as case report tabulations (CRTs) and analysis datasets. CRFs are the forms used by the clinical investigator to document the collected data. The CRTs are aggregate (i.e., data from multiple subjects grouped together) listings of all the data collected on the case report forms. CRFs and CRTs allow the Agency to perform an independent analysis of the study data. FDA routinely performs its own independent analyses of study data to assess the effectiveness and safety of investigational products. Analysis datasets are a subset of all data collected in the study and are the critical dataset to support the primary study analyses contained in the study report.

The animal studies are generally nonclinical pharmacology, toxicology, or biocompatability studies intended to investigate the actions of the investigational product in relation to its proposed therapeutic indication and intended clinical use.[4] Similar to clinical trials, applicants submit study reports that include pertinent tabulations of the collected and analyzed data. FDA reviewers use these data for safety and predictive assessments, which can include reanalysis of the submitted data.

Existing Federal regulations allow the voluntary submission of electronic records, including study data, to the Agency in lieu of paper records (see 21 CFR part 11). For many years, FDA Centers have requested that clinical study data be submitted electronically because paper CRTs are widely recognized as being highly inefficient to support analysis and review. The data in paper CRTs are not machine-readable and therefore cannot be easily analyzed using modern analytic software. Although submission of clinical study data in electronic format has become relatively routine at some Centers, these data are often not standardized. An important goal of

[4] In some circumstances, animal studies serve as efficacy studies (see 21 CFR 314.600 and 21 CFR 601.90).

this guidance is to increase the quantity of standardized electronic study data submitted to FDA for both clinical and nonclinical studies.

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 regulatory decisions.[5] As a result, FDA has publicly announced that it intends to propose a new Federal regulation that would require the submission of standardized electronic study data.[6] This guidance will support the new regulation once the rule is finalized.

Data that are not standardized diminish the Agency's ability to review the data efficiently, resulting in manual, labor-intensive processes and inherent inefficiencies in the review. They also limit the ability to automate some routine analyses. A simple example is an analysis of the distribution of the gender of subjects in a study. Gender can be represented in a study as male or female, M or F, or 0 or 1 (or 1 or 2; the possible representations are endless). If sponsors always expressed the gender of a subject in a study as M or F, then it is fairly trivial to create a computer program that can automatically count all occurrences of M and F and automatically generate a standard gender distribution report. If the next study uses 0 and 1 to represent gender, then the program is no longer useful. Considering that a typical study has many hundreds of data elements,[7] many with almost limitless ways they can be represented, the challenges in analyzing nonstandard data are quite evident.

As sponsors and applicants move toward standardized electronic study data submissions, there is a need to understand FDA's expectations for such data submissions. This guidance provides FDA's current thinking on the submission of standardized electronic study data.[8]

This guidance does not describe *what* data should be submitted. The content of a study data submission is largely determined by science and regulation, and often involves prior discussion with the appropriate review division to ensure that the content of the submission is adequate for its intended purpose. *What* to submit is outside the scope of this guidance; but having once decided what to submit, sponsors can consult this guidance to understand *how* to submit the data electronically, using data standards supported by the Agency.

To better understand why the Agency is now emphasizing the submission of standardized data for all studies, please refer to the Appendix.

III. GENERAL SUBMISSION CONSIDERATIONS

[5] Standardized data also facilitate data exchange and sharing (e.g., between a contract research organization and a sponsor).

[6] See http://www.reginfo.gov/public/do/eAgendaViewRule?ruleID=284747.

[7] By *data element*, we mean the smallest (or *atomic*) piece of information that is useful for analysis. Please see the Appendix for additional information.

[8] See section III.E for exceptions to standardized study data submissions.

Applicants and sponsors should refer to the FDA Study Data Standards Resources Web page[9] to obtain an up-to-date inventory of the various data standards supported[10] by CDER, CBER, and CDRH for standardized study data submissions.

To fully realize the advantages and benefits of data standardization, sponsors should consider during the planning phases of the study which data standards are applicable. This allows for streamlining data collection (e.g., data collection instruments and processes can be designed to collect study data in a standardized format or in a format that is amenable to standardization) and ensuring that all relevant data elements are collected in the proper format. Some sponsors choose to convert nonstandardized data to a standardized format. In general, we discourage this conversion approach unless it is the only option.[11] These two approaches (collecting standardized data vs. conversion of nonstandardized collected data) have somewhat different considerations that are discussed below.

A. Planning and Providing Standardized Study Data

For clinical and nonclinical studies, sponsors should describe in the IND or IDE their plan to submit standardized study data to the Agency. For an IND, the plan should be located in the general investigational plan.[12] For an IDE, the plan should be part of the IDE study protocol under the *Data Management Plan*. This study data standardization plan should contain a listing of the planned studies, or a description of the types of studies, along with the data standards that the sponsor intends to use. Where it is not possible or practical to use the FDA-recommended standards, the study data standardization plan should describe the studies or types of studies that will not be standardized, and describe why the use of FDA-recommended standards is not feasible. The inclusion of a study data standardization plan gives the Agency the opportunity to identify data standardization issues early in the development program. Sponsor feedback on feasible use of FDA-recommended standards may provide future opportunities to refine or develop data standards. The study data standardization plan should be updated in future communications with the Agency as the development program evolves.

When using a data standard, there may be occasional ambiguity resulting in more than one way to implement the data standard.[13] Instances in which a standard allows for more than one implementation should ideally be discussed with the appropriate review division prior to data submission. Submitters should ensure their data are fully compliant with the standard (see

[9] See http://www.fda.gov/ForIndustry/DataStandards/StudyDataStandards/default.htm.

[10] By *supported*, we mean that the FDA has established processes and technology infrastructure to support the receipt, processing, review, and archiving of study data using these standards.

[11] The conversion approach after the study is complete may disclose problems with data collection that could have been corrected earlier had standardized data been collected from the outset.

[12] See 21 CFR 312.23(a)(3) for a description of the general investigational plan. Note that a study data standardization plan is not a mandatory portion of the general investigational plan, but we recommend it be included in this section of the IND.

[13] For example, the CDISC SDTM Implementation Guide v. 3.1.2 describes the reference start date (RFSTDTC) as *usually* equivalent to date/time when subject was first exposed to study treatment. The word *usually* indicates other interpretations of the reference start date are possible.

section III.D, Data Validation). Any data that do not easily conform to the standard should be described in the study data submission.

At the time of a study data submission, submitters should describe in the cover letter how the study data standardization plan was implemented.[14] In addition, the submission should include a tabular listing of studies that contains a description of which data standards were used, including versions.[15]

A particular data standard does not require all data elements defined by the standard to be collected in any given study. For example, the Study Data Tabulation Model (SDTM)[16] classifies variables as required, expected, or permissible. As previously described, *what* is collected and submitted is the subject of science, regulation, and discussions with the review division. However, all data collected in a study should be submitted, and with the highest level of standardization possible.

Standardized electronic datasets should generally be accompanied by documentation typically referred to as a *Reviewer's Guide*. The Reviewer's Guide describes any special considerations or directions that will facilitate a reviewer's use of the datasets and can help the reviewer understand the relationships between the study report and the data. The Reviewer's Guide is an integral part of a standards-compliant study data submission. No additional guidance is available regarding the format and content of a Reviewer's Guide at this time.

B. Controlled Terminologies

The use of terminology standards, also known as controlled terminologies, is an important component of standardization and is a critical component of achieving semantically interoperable data exchange.[17] Terminology standards specify the key concepts that are represented as preferred terms, definitions, synonyms, codes, and code system. Terminology standards are maintained by external organizations (i.e., external to the submitter). Sponsor-defined custom terms are not considered controlled terminologies. Examples of controlled terminologies include:

- MedDRA (the Medical Dictionary for Regulatory Activities) for adverse events[18]
- The National Drug File — Reference Terminology for drug classifications[19]
- The International Standards Organization (ISO) 3166-1 alpha-3 standard for the countries of the world[20]

[14] This could include a simple statement saying that the standardization plan dated yyyy-mm-dd was executed faithfully, or that particular studies were not standardized according to the plan for stated reasons.
[15] For clinical studies to CBER and CDER, this information can be included in section 5.2 Tabular Listing of All Clinical Studies within the eCTD (electronic common technical document).
[16] See http://www.cdisc.org.
[17] See the Appendix for a detailed discussion of semantic interoperability.
[18] See http://www.meddramsso.com/.
[19] See http://ncit.nci.nih.gov/ncitbrowser/pages/vocabulary.jsf?dictionary=National%20Drug%20File%20-%20Reference%20Terminology.
[20] See http://www.fda.gov/ForIndustry/DataStandards/StructuredProductLabeling/ucm162567.htm.

- CDRH Event Problem Codes — used by CDRH to describe device problems and patient problems associated with device use[21]

The analysis of study data is greatly facilitated by the use of controlled terms for clinical or scientific concepts that have standard, predefined meanings and representations. The use of standard terminology for adverse events is perhaps the earliest example of using data standards for study data. *Myocardial infarction* and *heart attack* are both synonyms for the same clinical concept, and as such should be mapped to the same term in a standard dictionary. This facilitates an efficient analysis of events that are coded to the standard term. Because assigning standard terms from a dictionary to actual concepts is sometimes as much an art as a science, FDA expects submitters to provide in the electronic study data submission the actual verbatim terms that were collected on the case report form (the so called *raw* term), as well as the coded term, so that review staff can evaluate the standardization process.

The use of standard terminology is particularly useful when applied across studies, as this greatly facilitates appropriate integrated analyses that are stratified by study and related cross-study analyses (e.g., when greater power is needed to detect important safety signals). Cross-study comparisons and multi-study pooled analyses frequently provide critical information for regulatory decisions, such as confirmation of efficacy,[22] exposure-response relationships,[23] and population pharmacokinetics.[24]

The FDA Study Data Standards Resources Web page[25] contains the currently supported terminology standards at CBER, CDER, and CDRH. When planning a study (including the design of case report forms, data management systems, and statistical analysis plans), the sponsor should identify which FDA-supported standard terminologies to use for submission. In cases in which a sponsor does not intend to use an FDA-supported standard terminology, the sponsor should consult with and obtain concurrence from the appropriate review division.

If a sponsor identifies a concept for which no standard term exists, we recommend that the sponsor submit the concept to the appropriate terminology maintenance organization as early as possible to have a new term added to the standard dictionary. We consider this *good terminology management practice* for any organization. The creation of custom terms for a submission is discouraged (i.e., so called *extensible* code lists) as this does not support

---

[21] See http://www.fda.gov/MedicalDevices/Safety/ReportaProblem/EventProblemCodes/default.htm

[22] See the guidance for industry on *Providing Clinical Evidence of Effectiveness for Human Drugs and Biological Products*, http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/ucm072008.pdf. We update guidances periodically. To make sure you have the most recent version of a guidance, check the FDA Dugs guidance Web page at http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/default.htm.

[23] See the guidance for industry on *Exposure-Response Relationships — Study Design, Data Analysis, and Regulatory Applications*, http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/ucm072109.pdf.

[24] See the guidance for industry on *Population Pharmacokinetics*, http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/ucm072137.pdf.

[25] See http://www.fda.gov/ForIndustry/DataStandards/StudyDataStandards/default.htm.

semantically interoperable study data exchange. Terminology maintenance organizations generally have well-defined change control processes, and sponsors should allow sufficient time for a proposed term to be vetted, as it is desirable to have the term incorporated into the standard terminology before the data are submitted. If the use of custom terms cannot be avoided, the submitter should clearly identify and define the custom terms within the submission and use them consistently throughout the application. This is particularly important for primary and secondary clinical endpoints and patient reported outcomes used as endpoints.

If a submitter identifies an entire information domain[26] for which FDA has not endorsed a specific standard terminology, then the submitter may select a standard terminology to use. We recommend that the submitter discuss this selection with the review division in advance. The same terminology should then be used consistently throughout all relevant studies within the application.

A frequent question is how to handle multiple versions of a dictionary within a single application. We recognize that studies are done at different times, during which different versions of a dictionary may be the most current. We expect sponsors to use the most recent version of an FDA-supported dictionary available at the time of coding (recognizing that a single study will use a single version).[27] It is understandable and therefore acceptable to have different studies use different versions of the same dictionary within the same application. However, important pooled analyses of coded terms across multiple studies (e.g., a pooled adverse event analysis of all pivotal trials in an integrated summary of safety) should be conducted using the same version of a dictionary. This is the only way to ensure a truly consistent and coherent comparison of clinical and scientific concepts across multiple studies. Sponsors should clearly document which terminologies and which versions of terminologies are used for every study within a submission (see section III.A). Submitters should direct any specific questions regarding the use of controlled terminologies to the appropriate review division.

C. Standardization of Previously Collected Nonstandard Data

Clinical and nonclinical study data that were previously collected in a nonstandard format are not always easily amenable to complete standardization. Typically, a conversion to a standard format will map every data element as originally collected to a corresponding data element described in a standard. Some study data conversions will be straightforward and will result in complete data in a standardized format. In some instances, it may not be possible to represent a collected data element as a standard data element. In these cases, the submission should document why certain data elements could not be fully standardized or were otherwise not included in the standardized data submission. In cases where the data were collected on a CRF but not included in the converted datasets, the omitted data should be apparent on the annotated CRF. The tabular listing of studies in a submission should indicate which studies contained previously collected nonstandard data that were subsequently converted to standard format.

[26] By *information domain*, we mean a logical grouping of clinical or scientific concepts that are amenable to standardization (e.g., adverse event data, laboratory data, histopathology data, imaging data).

[27] That is to say, if a new version of a dictionary is released in the middle of the coding process for a given study, then the sponsor should continue to use the version that was most current when the coding process began.

D. Data Validation

For purposes of this guidance, data validation is a process that attempts to ensure that data are both compliant and useful. *Compliant* means the data conform to the applicable data standards. *Useful* means the ability of the data to support the intended use (e.g., regulatory review and analysis). Data validation is one method used to assess data quality.[28] Data validation relies on a set of validation rules that are used to check that the data conform to a minimum set of quality standards. Data validation attempts to identify early in the review common or serious data issues that may adversely affect use of the data. We recognize that data validation is an imperfect process, as it is impossible or impractical to define *a priori* all the relevant validation rules for any given submission. Sometimes serious deficiencies in the data are only evident through manual inspection of the data and may only become evident once the review is well under way. Often these are because of problems in data content (i.e., *what* was or was not submitted, or issues with the original source data), and not necessarily in *how* the data were standardized. We expect the data validation rules will undergo improvements as the Agency incrementally refines its definition of valid, high quality data for regulatory decisions.

Cross-study data validation for consistency is especially important for analyses supporting the confirmation of efficacy, exposure-response relationships, and population pharmacokinetics. For example, the terminology of the observational data elements (see section III.B) and the units of their values should be consistent across relevant studies.

FDA generally recognizes two types of validation rules:

Technical validation rules help ensure that the data conform to the data standards. For example, a technical validation rule for Clinical Data Interchange Standards Consortium (CDISC) Study Data Tabulation Model (SDTM) data would check that the value in the DOMAIN column of all datasets matches the name of the domain.

Business validation rules help ensure that the data will support business processes that rely on the data (i.e., support meaningful analysis of the data). For example, a business validation rule for a human study may require that each value for AGE fall in within a prespecified human physiologic range.

Once a data standard is defined, the technical validation rules are rather static. They are not expected to change substantially unless the standard itself changes. Business validation rules may evolve over time as new analysis requirements are identified and incorporated into data validation processes.

[28] Data quality is beyond the scope of this guidance. For purposes of this guidance, standardized data do not ensure quality data; however, standardized data do make it easier to assess some aspects of data quality by facilitating the automation of various data quality checks (e.g., completeness, reasonableness). Data standards can also improve the collection of quality data by facilitating the use of standard data quality checks in electronic data collection instruments.

Submitters can find the validation rules used at FDA on the Study Data Standards Resources Web page.[29] Submitters should validate their data before submission using the published validation rules and either correct any validation errors or explain in the submission why certain validation errors could not be corrected. The submission should state the version of the validation rules used.

Upon submission, FDA will conduct its own data validation, using the same version of the validation rules used by the sponsor. FDA will use the data validation results to inform review staff of potential problems in using the data, and to assess the usefulness of the rules. If necessary, FDA will report serious data validation errors to the submitter for correction. Please refer to the online validation rules for a description of data validation errors. The recommended presubmission validation step is intended to minimize the presence of serious validation errors at the time of submission.

E. Exceptions to Standardized Study Data Submissions

Data standards are evolving at different rates for varying study types, and they are not being adopted at the same rate throughout FDA. Accordingly, there may be some exceptions to FDA's general expectation that all study data be standardized for submission. The data standards catalog on the Study Data Standards Resources Web page describes which types of studies should be submitted in a standardized format, and which offices are currently accepting standardized study data submissions.

F. Meetings With FDA

Sponsors can use established FDA-sponsor meetings (e.g., pre-IND, pre-IDE, and end-of-phase 2) to raise data standardization issues (if any). Discussions about nonclinical study data standardization plans can be initiated at the pre-IND stage and should continue throughout development. Initial discussions about which data standards to use for clinical study data should take place as early as possible during drug development, especially for safety data, which is often less well planned for in advance, but should occur no later than the end of phase 2. In general, the premarketing application meeting is considered too late to initiate data standardization discussions.

Sponsors and applicants may submit technical questions related to data standards at any time to the technical support team identified by each Center (see the Study Data Standards Resources Web page for specific contact information). Submitters may also request a separate Type C meeting to discuss substantive data standardization issues. An example of such an issue might be a sponsor's desire to use a non-supported terminology. The request should include adequate information to identify the appropriate FDA staff necessary to discuss the proposed agenda items.

[29] See http://www.fda.gov/ForIndustry/DataStandards/StudyDataStandards/default.htm.

IV. ONLINE TECHNICAL RESOURCES

The FDA maintains online study data standards technical resources at http://www.fda.gov/ForIndustry/DataStandards/StudyDataStandards/default.htm, which includes the following:

- A data standards catalog that lists the various data standards (including versions of those standards) currently accepted at each Center
- Study data technical specifications document
- Study data validation rules
- Other Center-specific resources

Submitters should use the data standards listed on this Web page for their study data submissions. Data standards for study data can be expected to evolve over time; therefore, submitters are encouraged to visit the Web page regularly to obtain the latest information.

Each Center maintains a Data Standards Plan that outlines the management and adoption of data standards. As new standards are supported, the Agency will support older standards and versions of standards for at least 2 years after the addition of a new standard or version. FDA will update the Web page as needed to reflect changes in data standards and associated technical specifications.

V. ADDITIONAL SUPPORT

Sponsors/applicants with questions on how to implement the FDA-supported study data standards should contact and work with FDA technical staff. Contact information is provided on the Study Data Standards Resources Web page.[30] Sponsors may also arrange to submit sample data for a presubmission technical review. The technical staff also welcomes any additional feedback or comments regarding the information posted on the Web page.

[30] See http://www.fda.gov/ForIndustry/DataStandards/StudyDataStandards/default.htm.

APPENDIX – DATA STANDARDS AND INTEROPERABLE DATA EXCHANGE

This appendix provides information on why FDA is now emphasizing the submission of standardized data for both nonclinical and clinical studies. It also provides some of the guiding principles for the Agency's long-term study data standards management strategies. An important goal of standardizing study data submissions is to achieve an acceptable degree of *semantic interoperability* (discussed below). This appendix describes different types of interoperability and how data standards can support interoperable data exchange now and in the future.

At the most fundamental level, study data can be considered a collection of data elements and their relationships. A data element is the smallest (or *atomic*) piece of information that is useful for analysis (e.g., a systolic blood pressure measurement, a lab test result, a response to a question on a questionnaire).

A data value is by itself meaningless without additional information about the data (so called *metadata*). Metadata is often described as *data about data*. Metadata is structured information that describes, explains, or otherwise makes it easier to retrieve, use, or manage data.[31] For example, the number *44* itself is meaningless without an association with *Hematocrit*. Hematocrit in this example is metadata that further describes the data.

Just as it is important to standardize the representation of data (e.g., M and F for male and female, respectively), it is equally important to standardize the metadata. The expressions Hematocrit = 44; Hct = 44, or Hct Lab Test = 44 all convey the same information to a human, but an information system or analysis program will fail to recognize that they are equivalent because the metadata is not standardized. It is also important to standardize the definition of the metadata, so that the meaning of a Hematocrit is constant across studies and submissions.

In addition to standardizing the data and metadata, it is important to capture and represent relationships (also called associations) between data elements in a standard way. Relationships between data elements are critical to understand or interpret the data. Consider the following information collected on the same day for one subject in a study:

[31] Metadata is said to "give meaning to data" or to put data "in context." Although the term is now frequently used to refer to XML (extensible markup language) tags, there is nothing new about the concept of metadata. Data about a library book such as author, type of book, and the Library of Congress number, are metadata and were once maintained on index cards. SAS labels and formats are a rudimentary form of metadata, although they have not historically been referred to as metadata. Additional discussion about metadata is outside the scope of this guidance other than to stress the importance of standardized metadata for study data submissions.

Systolic Blood Pressure = 90 mmHg
Position = standing
Systolic Blood Pressure = 110 mmHg
Time = 10:23 a.m.
Time = 10:20 a.m.
Position = lying

When presented as a series of unrelated data elements, they cannot reliably be interpreted. Once the relationships are captured, as shown simply below using arrows:

Time = 10:20 a.m. ←→ Position = lying ←→ Systolic Blood Pressure = 110 mmHg
Time = 10:23 a.m. ←→ Position = standing ←→ Systolic Blood Pressure = 90 mmHg

the interpretation of a drop in systolic blood pressure of 20 mmHg while standing, and therefore the presence of clinical orthostatic hypotension, is possible. Standardizing study data therefore involves standardizing the data, metadata, and the representation of relationships.

With these fundamental concepts of data standardization in mind, data standards can be considered in the context of interoperable data exchange.

A. Interoperability

Much has been written about interoperability, with many available definitions and interpretations within the health informatics community. In August 2006, the President signed an Executive Order mandating the Federal Government use of interoperable data standards for health information exchange.[32] Although this order was directed at Federal agencies that administer health care programs (and therefore not FDA), it is relevant to this guidance because it defined interoperability:

"Interoperability" means the ability to communicate and exchange data accurately, effectively, securely, and consistently with different information technology systems, software applications, and networks in various settings, and exchange data such that clinical or operational purpose and meaning of the data are preserved and unaltered.

Achieving interoperable study data exchange between submitters and FDA is not an all-or-nothing proposition. Interoperability represents a continuum, with higher degrees of data standardization resulting in greater interoperability, which in turn makes the data more useful and increasingly capable of supporting efficient processes and analyses by the data recipient. It is therefore useful to understand the degree of interoperability that is desirable for standardized study data submissions.

[32] See http://www.cga.ct.gov/2006/rpt/2006-R-0603.htm.

In 2007, the Electronic Health Record Interoperability Work Group within Health Level Seven issued a white paper that characterized the different types of interoperability based on an analysis of how the term was being defined and used in actual practice.[33] Three hierarchical types of interoperability were identified: technical, semantic, and process interoperability. A review of these three types provides insight into the desired level of interoperability for standardized study data submissions.

Technical interoperability describes the lowest level of interoperability whereby two different systems or organizations exchange data so that the data are useful. There is nothing that defines how useful. The focus of technical interoperability is on the conveyance of data, not on its meaning. Technical interoperability supports the exchange of information that can be used by a person but not necessarily processed further. When applied to study data, a simple exchange of nonstandardized data using an agreed-upon file format for data exchange (e.g., SAS transport file) is an example of technical interoperability.

Semantic interoperability describes the ability of information shared by systems to be understood, so that nonnumeric data can be processed by the receiving system. Semantic interoperability is a multi-level concept with the degree of semantic interoperability dependent on the level of agreement on data content terminology and other factors. With greater degrees of semantic interoperability, less human manual processing is required, thereby decreasing errors and inefficiencies in data analysis. The use of controlled terminologies and consistently defined metadata support semantic interoperability.

Process interoperability is an emerging concept that has been identified as a requirement for successful system implementation into actual work settings. Simply put, it involves the ability of a system to provide the right data to the right entity at the right point in a business process.

In a regulatory setting, an example of process interoperability is the ability to quickly and automatically identify and provide all the necessary information to produce an expedited adverse event report in a clinical trial upon the occurrence of a serious and unexpected adverse event. The timely submission of this information is required by regulation to support FDA's mandate to safeguard patient safety during a clinical trial. Process interoperability becomes important when particular data are necessary to support time-dependent processes.

Because the vast majority of study data are submitted after the study is complete, achieving process interoperability for study data submissions in a regulatory setting is relatively unimportant, at least for the foreseeable future. It is reasonable to conclude that it is most desirable to achieve *semantic interoperability* in standardized study data submissions.

[33] See Coming to Terms: Scoping Interoperability for Health Care http://www.hln.com/assets/pdf/Coming-to-Terms-February-2007.pdf.

B. Types of Data Standards

There are various types of data standards, and all have a role in achieving semantically interoperable study data exchange. The following describe working definitions that are not necessarily mutually exclusive.

A **file format standard** specifies a particular way that information is encoded in a computer file. Specifications for a format permit the file to be written according to a standard, opened for use or alteration, and written back to a storage medium for later access. Some file formats in widespread use are proprietary, others are open source. Examples of file format standards supported at FDA include Adobe Acrobat Portable Document (.pdf), SAS Transport File format (.xpt), text files (.txt), and Extensible Markup Language (.xml). The use of a file format standard for study data exchange supports technical interoperability, but by itself is often insufficient for semantic interoperability.

An **exchange standard** describes a standard way of exchanging data between computer systems. Exchange standards may describe the data elements and relationships necessary to achieve the unambiguous exchange of information between disparate information systems. Examples include the Health Level 7 (HL7) Structured Product Labeling (SPL), Individual Case Safety Report (ICSR), and ECG Waveform standards. The use of an exchange standard is necessary but often not sufficient to achieve semantically interoperable data exchange. Often the exchange standard should be combined with standard terminologies to achieve an acceptable degree of semantic interoperability.

A **terminology standard** is the group of specifications for commonly agreed-upon vocabulary to be used in another data standard or family of data standards. Terminology standards specify the key concepts that are represented as preferred terms, definitions, synonyms, codes, and the code system that must be used to ensure a common understanding of a given data standard. Examples include the ISO 3166-1 alpha-3 standard for representing the countries of the world, the FDA/USP Substance Registration System's Unique Ingredient Identifier (UNII) for representing chemical and biological substances, the Department of Veterans Affairs' National Drug File – Reference Terminology for classifying drugs. Terminology standards, when used within exchange and file format standards, combine to provide a useful degree of semantically interoperable data exchange.

An **analysis standard** describes a standard presentation of the data intended to support analysis. It includes extraction, transformation, and derivations of the original data. An example is the CDISC Analysis Data Model (ADaM) standard.

As a practical matter, most of the existing data standards do not have one-to-one correspondence with these types of data standards. For example, the CDISC Study Data Tabulation Model (SDTM) can be used as an exchange standard (when combined with the SAS transport file format standard) and as an analysis standard to support simple analyses, such as demographics or adverse events.

Those familiar with data standards often describe other types of standards: data organization standards, content standards (such as disease/domain specific standards (e.g., for Alzheimer's disease or Diabetes Mellitus)), and data presentation standards. We find these are all variations on the previously described data standard types.

In summary, the goal of standardizing study data is to make the data more useful and to support semantically interoperable data exchange between submitters and the FDA. This, in turn, will support more efficient analytic processes at the Agency. The combined use of various types of data standards is necessary to achieve an acceptable degree of semantic interoperability.