



Common Data Elements Are Not Enough

The fashion these days is to standardize at the individual data element level. Thus, when standards are developed, they tend to focus more on characteristics of the data element alone and less on its relationship with other data elements. While this may initially seem to be the simplest approach, it risks omitting information that later users may need to interpret the data correctly, especially when the data are repurposed, such as being pooled for meta-analyses or data mining. If “quality data” are defined as “data that are fit for their intended uses,”¹ then this threatens the quality of the data and the conclusions drawn from them.



Table of Contents

- Pg. 1...Why Standardizing Data Elements is Not Enough
- Pg. 2...Fun Pharma Facts.
- Pg. 7...Consultants' Corner
- Pg. 8...Fan Forum

Data Elements Libraries

Many organizations are developing common data element libraries, from CDISC² to many parts of the US National Institutes of Health³ to medical associations such as the American Heart Association⁴. They differ in their uses of the data, e.g., for patient care vs. clinical research, but there are basic traits that most include in their definitions (see Table 1). Many, including those referenced above, are robust and resolve many of the issues they were designed to address,

such as different variable names and code lists, conflicting types (e.g., character vs. numeric) and so forth, and are making data sharing far easier and more reliable.

It is logical to organize standards by data element; from a data collection point of view, it is the smallest independent unit that cannot be further subdivided, and it can be grouped in different ways to form case report forms, health care charts, etc.

This approach has its drawbacks however. Many elements are not independent, and either need or are supportive of other elements, without which they lose their meaning. Some relationships are, at least to the experienced user, self-evident and many think they do not need to be defined, but other relationships are less so. Where and how to define them is a challenge, as they belong to all mem-

Table 1. Data Element Characteristics that are Commonly Standardized

Characteristic	Definition
Data Element Name	Brief human-language label for the question asked or info solicited
Variable name	Electronic name of the data element (e.g., in a database)
Definition	Further description ensuring unambiguous meaning
Type	Format of response (e.g., character, numeric, date)
Code List	A set of pre-defined answers; checklist; aka controlled terminology
Completion Instructions	Directions on how to respond to the question
Associated Domain	The category of data to which the data element belongs, e.g., AEs, Concomitant Medications, Demographics, etc.

(Continued on page 3)

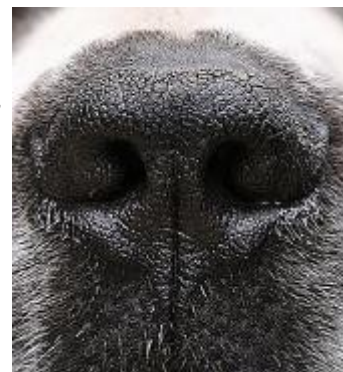
Fun Pharma Facts

Did you know?

Cancer leaves a distinct chemical signature on the breath.

Researchers at the Pine Street Foundation in California have trained dogs to smell various kinds of cancer on a patient's breath. The dogs' diagnoses were 88 to 99 percent accurate, even with other factors like diet and smoking taken into account. Even with such a high success rate, you will most likely not see dogs running around a doctor's office or hospital randomly administering the "sniff test" on patients, as amusing as that might be.

Instead, scientists have been developing machines that can diagnose diseases by sampling a patient's breath. Certain diseases cause the body to produce chemicals that we normally don't produce, or in different amounts than a healthy person would produce. Some of these chemicals find their way into the air we exhale, which means they can be smelled.



The challenge is to figure out what chemicals to look for. That's actually tougher than it sounds, because the smell of cancer is probably made up of many different chemicals. What's more, each type of cancer may have a unique chemical signature. It will take much testing and experimentation (possibly involving humans, dogs, and machines) to find just the right formula.

If the testing works, it could make screening for deadly diseases cheaper and less invasive.

Source: 10, February 2006, Podcast, www.scienceupdate.com

Your Response Has Been Overwhelming!
We're Extending Our Sale Until Feb. 15!

Save 30% on All Kestrel
On-Demand Courses
During the Month of January!

Buy Now and Take the Course at Your Convenience!
To Cash in on the Savings, Use Coupon Code **KST030**

KestrelConsulting.trainingcampus.net

To access Kestrel's Courses, please find the Product Catalogue button under the Kestrel Header. Then you may search the course catalogue and make your purchase; pricing depends on whether you work in academia or Industry.

Receive an extra Price Break if you work in Academia or Government!

Kestrel's CYPHER, Mapping Your Standards To CDASH

Harness our Expertise To Save you Time and Money



Inquiries@kestrelconsultants.com



Why Standardizing Data Elements is Not Enough, cont.

(Continued from page 1)



bers of the relationship, but it is not good practice to duplicate information, and some relationships may not be applicable in all studies. It is also increasingly common for data to have many lives, from capture for a clinical study, to inclusion in a regulatory submission, to data repositories, to supporting clinical care, and many other potential uses that have not even been envisioned yet. Even supposedly self-evident relationships may not be so obvious in other contexts.

Because the data will outlive their initial purpose, users should understand the assumptions and constraints that influenced the data elements' definitions, or they may use the data inappropriately. Even if the data elements library accompanies the data and the elements were used exactly as defined in the library, neither of which may be true, libraries often do not define these relationships, and it may not be appropriate for them to do so. This is especially critical as standards are often seen as a good way to help inexperienced colleagues to collect the right data in the right way, and without guidance they may not do so appropriately. Following are some recent cases that illustrate these points, some of which are obvious but others that are not.

Cases

Measurements and Units of Measure

This is the most obvious example. Measurements have no meaning if the unit of measure is not defined. For example, a weight of 25 is not useful unless we know if it is in kilos or pounds. Both are reasonable pediatric values, but can have different implications depending upon the age or condition of the subject. Many organizations preprint or "paint" the units on the CRF because the form filler either just knows what unit to use or the protocol defines it. When the data are entered into a data-



(Continued on page 4)

Why Standardizing Data Elements is Not Enough, cont.

(Continued from page 3)

base or electronic CRF (eCRF), the units may be placed in the label of the measurement or may not be captured electronically at all, assuming that since the protocol defined the unit it can only have one value and therefore does not have to be “databased.” If these data are later pooled with data from another organization, one can no longer assume that the units are the same and weights become unusable.

Adverse Event and Severity of Event

There are two potential challenges here. The first is that Severity of Event only has meaning in relation to the adverse event (AE). This may seem obvious, but when specific AEs are assessed they are often pre-printed on the CRF, and because the focus is on the study rather than data pooling, it is assumed that everyone knows what the AE was. Indeed, if the AE term is not “databased,” all the AE data become unusable, not just the severity. More generally, any element that describes or modifies another element may not be useable unless the modified element is included.

The second challenge, while not directly about combinations of elements, is also a potential quality issue. There are typically 2 code lists (or sets of controlled terminology) used with the Severity field: *Mild*, *Moderate* and *Severe* (for most AEs) and the same three plus *Life-Threatening* and *Fatal* for oncology studies. It should be very clear which one is used in each study, and this information should always accompany the data. This is because if only *Mild*, *Moderate* and *Severe* are present in the database, it is unknown if *Life-Threatening* and *Fatal* were not on the CRF or were not observed in the study. This is true for all cases where subsets of code lists are used.



Reference Ranges (aka Normal Ranges)

Reference ranges are elements that define the expected values of a specific response. The most common are the upper and lower limits for laboratory tests. For example, it is not possible to interpret the significance of a urine glucose value of *20 mg/dL* in the absence of the expected range (*0 - 15 mg/dL*). While ranges for many tests can be found in textbooks or online, these may not be applicable to the subject population. Although it requires more storage space, it is best to include the relevant reference range on every record in a lab dataset, and this requirement should be specified in the standard data element library.



(Continued on page 5)

Why Standardizing Data Elements is Not Enough, cont.

(Continued from page 4)

Quality Assessments

A set of CRFs captured information about images, and the following question was included:

Indicate the quality of the image:

Exemplary quality

Adequate quality

Limited quality

Not adequate quality



The quality of the image affects its interpretability and the confidence in the result, In this case, the quality of the image is judged in the context of the study’s requirements, i.e., “Is it good enough for its current purpose?”

In the case of 3-D images, such as MRIs, images consist of a series of “slices,” each of which is a 2-D image that is stacked with other slices to build a 3-D picture. It is possible for one part of a 3-D image to be good quality and another to be poor quality, or some slices to be fine and others to be unusable for a given purpose. A single question asking about the quality of the image does not distinguish the quality of the entire image vs. just one or a few slices. Whether this is an issue depends upon how the question is used. If in one study it assessed the entire 3-D image while in a second study it assessed sets of at least 15 slices, a pooled dataset of all *Exemplary* images will not be consistent or comparable. On the other hand, if the quality question is used only to determine if an image interpretation should be present, then there is no issue. This assumes, however, that the intent of the question is clear to users later in the data lifecycle, and today this is an unsolved challenge.

Table 2. Data Elements Present in Data Elements Library

#	Data Element	Code List
1	Reason test was not completed	Equipment failure/error
		Medical reason
		Other
		Participant death
		Participant refusal
		Participant withdrew
		Scheduling problem
		Unknown
2	Other reason test was not completed	(open text field)
3	Medical reason test was not completed	Abnormal laboratory level
		Adverse Event
		Claustrophobia
		Injection complication
		Progressive disease

Overlapping Elements

In a data elements library reviewed recently, the three elements in Table 2 were present. Each question by itself is fine, but as is often the case, there was no indication of what fields were intended to be used or not used

(Continued on page 6)

Why Standardizing Data Elements is Not Enough, cont.

(Continued from page 5)

together. Question 1 asks for a general reason why the test was not completed, and one of the responses is *Medical Reason*. Question 2 provides a place to specify a reason if it was not included in the list. Question 3 asks for the specific medical reason. If all three are used together, they provide two places to capture that there was a medical reason, which would need to be cross-checked for consistency. If the reason was not medical, Question 3 would be blank as there is no way to indicate that it was not a medical reason. If the reason was a medical one, but not listed in Question 3, it is quite likely that a form filler might check *Medical Reason* for Question 1, specify the reason in Question 2, and leave Question 3 blank.

This issue could be resolved either by not using the three elements together, or by laying out the CRF such that the Medical Reason list from Question 3 was next to the *Medical Reason* response in Question 1, although this works better for paper CRFs than it does for electronic ones. The data elements library did not provide either layout or element combination information, which could lead to the same data being captured inconsistently.



Conclusions



In order for data to be high quality, users (including potential future users) must understand how to use the data appropriately. These examples all demonstrate that having standard data elements is not sufficient to ensure high quality data, and that users must also understand, among many other things, the relationships that elements have to each other and how to use them together. Many will say that this is obvious, and that people who do not understand it should not be designing clinical trials data. This may or may not be true, but what is obvious to one person is not necessarily obvious to another, especially in different therapy areas or study designs.

Sometimes including these relationships in data libraries may be appropriate, but in many cases it is not, either because they are very study- or institution-specific or for some other reason. Also, these libraries are not typically stored with the data and may not be available to later users. Instead, there should be a way to associate

(Continued on page 8)

Consultant's Corner

In every issue, we highlight members of our consulting community by listing their company's name, contact information and specialty. The goal is to present our readers with a resource to utilize if and when a consultant's expertise is needed.

Name: Deborah Salerno, PhD
 Company: Salerno Scientific
 Email: Deborah@SalernoScientific.com
 Phone: (734) 662-1572
 Website: <http://salernoscientific.com>
 Services: Medical writing, consulting, workshops




Name: Richard McLain, MS
 Company: PFP Statistical Consulting, LLC
 Email: richard.mclain@ameritech.net
 Phone: 734-266-0100
 Services: STATISTICAL SUPPORT: Over 20 years experience working for a large pharmaceutical company. Providing support with study design, sample size calculation, protocol development, statistical analysis plans, interim analyses, DSMB reports and final analyses



Name: Anne B Giordani PhD, ELS
 Company: Principal, Anne B Giordani PhD, LLC
 Email: anne.giordani@comcast.net
 Phone: 734-665-2256 & 734-657-1097 (cell)
 Services: Medical, Scientific, and Regulatory Writing and Editing




Name: Lori Weaver and Lynne Welling 
 Company: LEAD Clinical Research, LLC
 Email: LoriatLEAD@sbcglobal.net
LynneatLEAD@charter.com
 Phone: 734-649-1963 and 586-202-7574
 Website: www.LeadCR.com
 Services: Consulting, Project and Study Management for clinical trials (Phase 1-4), inspection readiness, SOP development, SAE processing



Name: Charles Schmidt
 Address: Rio de Janeiro St 212, 15 floor, Sao Paulo – SP – Brazil (ZIP CODE 01240-010)
 Phones: +55-11 9970 2473 or +55-11 3667 1974
 Services for Brazil: Clinical Operation, Regulatory, Safety, Data Management and Statistics, Medical Writer, and Central Lab



Name: Robert Musterer, MBA 
 Company: ER Squared, Inc.
 Email: RMusterer@er2inc.com
 Phone: 203-974-3296
 Website: www.er2inc.com
 Services: consulting and auditing services in the areas of clinical data management, electronic data capture (EDC), e-clinical, data repositories, and Outsourcing



*Why Standardizing Data Elements is Not Enough, cont.**(Continued from page 6)*

these design and handling rules with the data as it progresses through its lifecycle, and they should be easy to review and apply. There is no such structure currently, and for the data repositories of the future to be robust and their data used appropriately, this issue will have to be addressed.

End Notes

¹ Institute of Medicine. *Assuring Data Quality and Validity in Clinical Trials for Regulatory Decision Making*. Jonathan R. Davis, Vivian P. Nolan, Janet Woodcock, and Ronald W. Estabrook, Editors. National Academy Press, c. 1999

² Clinical Data Interchange Standards Consortium. www.cdisc.org

³ National Institutes of Health: These are some institutes that have published CDE information

National Cancer Institute Data Standards caBIG: <https://caBIG.nci.nih.gov>

National Institute of Neurological Disorders and Stroke:
http://www.ninds.nih.gov/research/clinical_research/toolkit/common_data_elements.htm

Office of Rare Diseases Research: <http://www.grdr.info/index.php/common-data-elements>

⁴ American Heart Association: <http://circ.ahajournals.org/content/112/12/1888.full>

**Fan Forum!**

"Working with Kit and the Kestrel staff has been tremendously beneficial to our team. Kit has helped us develop industry-based standards and apply them to the unique needs of an academic medical center. Her training modules are outstanding and have allowed us to expand our knowledge base tremendously!"

Rick Ittenbach, Associate Professor of Pediatrics, Biostatistics at Cincinnati Children's Hospital Medical Center

