JOHNS HOPKINS
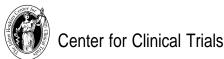U N I V E R S I T Y

Center for Clinical Trials

Department of Biostatistics
Department of Epidemiology
Department of International Health

Department of Medicine
Department of Ophthalmology
Oncology Center

(Monday 7:04am) 12 June 2000

Memorandum

To: Center for Clinical Trials faculty and staff

Fr: Curt Meinert

Re: Data entry good practice policies and procedures (GPPP)

Definitions

centralized data entry *n* - A form of **data entry** in which **data**, generated and recorded at different **sites** are sent to a **central** facility for conversion to machine-readable form by **coding** and **keying** for **storage** in a **database**. ant: **distributed data entry** rt: **centralized data analysis**

centralized data system *n* - A **data system** that is established and maintained at a central site (eg, the **data coordinating center**) in a **multicenter** structure; **site** is responsible for performing defined tasks in relation to maintaining that system. ant: **distributed data system**

centralized database *n* - A **database** residing and maintained at a **central** facility, eg, at the **data coordinating center** in a **multicenter trial**. ant: **distributed database**

data *n* - [L, pl of **datum**] Factual information, such as **measurements**, **observations**, or **statistics**, used as a basis for reasoning, discussion, or calculation. In **trials**, the term usually refers to information collected and recorded on **patients** considered for **enrollment** or actually enrolled. rt: **observed data**, **raw data** *Usage note*: Plural, but also singular, especially in reference to a set or mass of data, as in *data is collected at weekly intervals*. Though use in the singular form is increasingly common, such usage is not recommended.

data audit *n* - 1. The **comparison** of **data** in a **source document** with those in a **secondary document** as a means of checking for discrepancies. 2. A comparison of data in a **study form** with those recorded in a **medical chart** for discrepancies. 3. A comparison of specified **data elements** on an **original study form** with the corresponding elements in a **record** or **electronic file** produced from the original for discrepancies.

data bank *n* - 1. A collection of **computer** readable **databases** or **programs**. 2. A place for storing **data**; **archive**; **repository**.

data center *n* - 1. A **center** in a **study** structure that is responsible for receiving, **editing**, processing, **analyzing**, and storing **study data**, but that has few if any of the other

---

responsibilities assumed by a **data coordinating center** or **coordinating center**. 2. **data coordinating center** rt: **biostatistical center**, **statistical center**

data classification *n* - The act or process of arranging **data** into **classes** or **categories**, eg, as needed for preparing a **frequency distribution**.

data coding *n* - A process involving the **assignment** and use of numeric and letter codes to denote **categories**, **classes**, or patterns of information represented on a **data form** or in a **record**, eg, the use of the letters M and F to denote the gender of a **study participant** or the assignment of number codes to the different wave forms on an electrocardiogram. rt: **data reduction**, **data transcription**, **data transformation**

data collection form *n* - A **data form** used for **data collection**.

data collection protocol *n* - 1. The **protocol** for **data collection**. 2. That portion of the **study protocol** concerned with data collection, especially that portion containing details of the **data collection schedule**, data collection procedures, and data to be collected. 3. The **data collection schedule** and associated **data collection forms**. rt: **study protocol**, **treatment protocol**

data collection schedule *n* - 1. The designated **time points** at which **data** will be collected in a **study**. The **schedule** may be specified by calendar time (as in a **trial** requiring **data collection** on all **study participants** every January and July) or relative to some starting point (as in a trial where data for each person are collected at **randomization** and at 7 and 14 days following randomization). 2. **data collection form** (not recommended usage)

data collection site *n* - 1. A **site** responsible for **data collection**, eg, a **clinic** in a **clinical trial** or a site that mails and receives forms in a **followup study** done by mail. 2. **data generation site** (defn 1) *Usage note*: Distinguish between data collection and data generation sites. A data collection site is, by definition, a data generation site, but a data generation site need not be a data collection site, eg, a **reading center** responsible for reading eye fundus photographs taken at a data collection site.

data collection visit *n* - 1. Any **visit** by a **study candidate** or **study participant** to a **study center** for **data collection**. 2. Any **visit** by study personnel to the home or place of residence of a **study candidate** or **study participant** for data collection.

data coordinating center (DCC) *n* - 1. A **center** in a **multicenter study** structure that has the duties of a **data center** (defn 1), as well as duties for coordination of activities related to **data collection**. 2. **coordinating center** (defn 1) *Usage note*: The modifier, **data**, is used to emphasize the data functions of the center, especially in settings involving two or more centers with coordination responsibilities (as in a structure with a **data coordinating center** and **treatment coordinating center**).

data coordinator *n* - 1. An individual in the **data center**, **data coordinating center**, or **coordinating center** responsible for coordinating the receipt of data from **study centers** and for communicating with them regarding data flow. 2. **study clinic coordinator** (defn 1 or 2)

data editing *v* - 1. The process of reviewing **data** for the purpose of detecting deficiencies or **errors** in the way they are collected or recorded. 2. The actual process of detecting deficient or erroneous values on completed **data forms**. rt: **data query**

data editing *v* - 1. The process of reviewing **data** for the purpose of detecting deficiencies or **errors** in the way they are collected or recorded. 2. The actual process of detecting deficient or erroneous values on completed **data forms**. rt: **data query**

data element *n* - A unit of **data** recorded or to be recorded for a specific **data item** or **data field**; **datum**.

data entry *n* - 1. The process of entering **data**, **raw** or edited, into some **database** for storage and subsequent use; typically done by keying responses to queries appearing on a **computer terminal** or by transcribing information contained on a **data form** to produce a mechanically or electronically readable **database**; **data keying**. 2. The process of completing a **data form**. rt: **centralized data entry**, **distributed data entry**, **double data entry**

data field *n* - A space on a **data form** or in an **electronic record** designated to contain, or that actually contains, alphabetic and/or numeric characters of information recorded in response to a specific **data item** on the form.

data file *n* - A collection of **data records**, as contained on paper forms or in electronic **records**, arrayed or organized in some fashion.

data form *n* - 1. A collection of **data items** as contained on or in a **data record**. 2. A **form** containing **data**. 3. A form, **paper** or **electronic**, for **data collection**. rt: **case report form**

data freeze *n* - **Data** held in a fixed state, especially such a state imposed on an active **database** or **data file** in order to complete some task requiring a stable, nonchanging, database or data file (eg, as required for preparation of a **treatment effects monitoring report**). rt: **data snapshot**

data freeze *v* - The act or process of **freezing** an active **database** or **file** in order to hold in a fixed state, eg, as required for preparation of a **treatment effects monitoring report**.

data generation *n* - 1. **data collection** (defn 1) 2. The **generation** of **data** from specimens, documents, or materials collected by others, eg, data generated from blood samples received at a **laboratory** or from readings of fundus photographs made at a **reading center**.

data generation site *n* - 1. A **site** that generates **data** from existing **records**, eg, a **reading center**. 2. **data collection site** *Usage note*: See note for **data collection site**.

data item *n* - 1. A written **request**, usually in the form of an instruction, statement, or question, and associated **data field** for **response**, especially such a request and field as contained in a **data form**. 2. The response to such a request. 3. **data field**

data keying *n* - A process involving use of a keyboard and defined key strokes to represent observed **data**; **data entry**.

data management *n* - A constellation of activities, typically performed by a **data center**, **data coordinating center**, or **coordinating center**, related to receiving, editing, updating, storing, and retrieving **data** for some specified task or function.

data manager *n* - One who manages **data** or who is involved in **data management**.

data monitoring *v* - 1. **Monitoring** related to the process of **data collection**. 2. Monitoring related to the detection of problems in the execution of a **study** (**performance monitoring**) or for detecting **treatment effects** (**treatment monitoring**). rt: **data and safety monitoring**, **performance monitoring**, **safety monitoring**, **treatment effects monitoring**

data monitoring committee (DMC) *n* - A **committee** with defined responsibilities for **data monitoring**, eg, as required in **performance** or **treatment effects monitoring**. syn: **data and safety monitoring committee**, **performance monitoring committee**, **safety monitoring committee**, **treatment effects monitoring committee**, **treatment effects and performance monitoring committee**, **treatment effects monitoring and analysis committee** *Usage note*: Not recommended because of ambiguities regarding domain of responsibility. Use a more specific term denoting the area of responsibility of the committee, eg, **treatment effects monitoring committee** when the committee has responsibility for **treatment monitoring**.

data monitoring report *n* - 1. A **report** of some aspect of **data monitoring**. 2. A **report**, containing summary **tabulations** and **analyses**, prepared in relation to a data monitoring activity; such a report as seen by a **data monitoring committee**; **performance monitoring report**; **treatment effects monitoring committee**.

data processing *n* - 1. The constellation of activities related to **inventorying** and entering **data**. 2. Those activities performed subsequent to **data entry** in relation to **editing**, **updating**, and **analyzing**; especially activities performed in relation to the creation and maintenance of an **electronic database**.

data purge *n* - The removal of specified **data elements** from a **dataset** because of known or suspected deficiencies; eg, data known to have been or suspected of having been fabricated.

data query *n* - A **query** regarding a **data element** or **item**. rt: **data editing**, **edit query**

data record *n* - A collection of data **items**, as contained in a **paper** or **electronic form**, treated as a **unit** for some defined purpose or function.

data reduction *n* - The process of condensing **data**, by codification, grouping, summarization, and other means so as to make them more amenable to storage and processing.

data simple *adj* - Of, relating to, or characterized by **data collection** procedures and requirements that are **simple** (defn 2).

data simple trial *n* - A **trial** having a **parallel treatment design** and **simple** (defn 2) **data collection** procedures, especially one in which essential **baseline** and **followup data** are recorded on a few short, simple forms and having death or some other **clinical event** that can be reliably observed without use of elaborate diagnostic tests or procedures, as the **outcome** of sole or **primary** interest. rt: **big trial**, **simple trial**, **big and simple trial** *Usage note*: See **simple trial**.

data snapshot *n* - A **snapshot** of an active **database** or **file**; generally not preceded by any special effort to ensure **clean data**. rt: **data freeze**

data system *n* - A collection of interrelated procedures and routines (performed by hand or **computer**) that are used in the creation and maintenance of a **database** or some data-related function, such as **data analysis**. rt: **centralized data system**, **distributed data system**

data transcription *n* - 1. **Transcription** of **data** onto a **data collection form**. 2. Transcription of data already recorded on a data collection form to some other form, document, or list; not to be confused with **data coding** or **data editing**.

data transformation *n* - 1. A one-to-one mapping of **data** from one **measurement scale** to another; typically performed to simplify calculations for **data analysis** or to provide data having distributional or other properties better suited than the **original** to some form of display or statistical analysis. 2. Any form of **data reduction** or simplification, regardless of whether or not a one-to-one mapping, eg, the distillation of multiple **outcome measures** into a single **composite outcome**. 3. **data coding** rt: **inverse sine transformation**, **log transformation**, **square root transformation**, **standard score**, **T-score**, **Z-transformation**, **Z-score**

database *n* - 1. A collection of related **data** or related **data files**, organized for ease and efficiency of use, especially in relation to activities such as those involving retrieving or updating specific files or **data fields**. 2. **analysis database** 3. **electronic database** rt: **centralized database**, **distributed database**, **study database**

database management system *n* - A collection of structured procedures and routines, usually involving **computers** and **programs**, used for entering, sorting, editing, storing, and retrieving **data**.

dataset *n* - 1. A collection of **information** organized and arrayed in some way for convenience of **access** and use. 2. **electronic dataset** *Usage note*: Subject to varying uses, often used in references to **electronic database** but may refer to collections of information not amenable to electronic processing or manipulation. In the setting of **trials**, often used in references to collections of **data** on defined **populations** observed at one or more **time points** over the course of **observation**, but may also refer to such collections for a single **variable**. rt: **database**

dependent double data entry *n* - **Double data entry** with both entries performed by the same person at the same time. rt: **independent double data entry**, **quasi-independent double data entry**

direct data entry *n* - 1. **on-line data entry** 2. **Data entry** done from **paper forms** completed during the **data collection** process without any intermediate **coding** or **transcription**. 3. Data entry done at the **generation site**. ant: **indirect data entry**

distributed data entry *n* - [**multicenter study**] **Data entry** performed at the **site** of **collection** or **generation**. syn: **remote data entry** ant: **centralized data entry**

distributed data system *n* - A **data system** consisting of component parts that are established and maintained at individual **data collection** or **generation sites** in a **multicenter** structure and where those sites perform functions similar to those performed at the **data coordinating center** for creating and maintaining a **centralized data system**. ant: **centralized data entry**

distributed database *n* - A **database** assembled from component parts residing at different institutions or at different locations in the same institution, eg, a database comprised of **data** generated and keyed at the various **data generation sites** in a **multicenter study**. ant: **centralized database**

electronic form *n* - A **form** that is or may be displayed electronically (eg, on a **video screen**), especially one that is or may be completed using such displays and associated electronic devices. ant: **paper form**

independent double data entry *n* - **Data entry** in which data are entered (keyed) by two different people working independently of one another; as distinct from **dependent double data entry**. rt: **quasi-independent double data entry**

indirect data entry *n* - 1. **Data entry** performed using code sheets or other **documents** produced from the original **study forms** and documents. 2. Data entry not performed at the **data**

generation site. ant: **direct data entry**

linkage error *n* - An **error** involving improper linkage of **data**, eg, a case where **baseline data** for one person are linked with **followup data** for another person.

paper form *n* - A **form** printed or completed on **paper**; as opposed to an **electronic form**. ant: **electronic form**

quasi-independent double data entry *n* - **Double data entry** with both entries performed by the same person but at different times, eg, entered on one day and reentered on another day. rt: **dependent double data entry**, **independent double data entry**

remote data entry *n* - [**multicenter study**] **Data entry** performed at **collection** or **generation sites** and transmitted to a central site by modem or other means for **analysis** and storage. rt: **distributed data entry**

what you see is what you get (WYSIWYG) - [**computers**] Jargon used to indicate that the appearance of a document displayed on the **video screen** of a **computer** will be the same as the printed version of that document; generally considered to be a desirable quality of word processing packages.

P&P 1: Design data collection forms for keying directly from the forms, without transcription or additional coding.

P&P 2: Design each data collection form to be:
• Independent of all other data collection forms
• Amenable to being opened and closed on the day of use
• Ready for immediate data entry when closed
Comment
  See *Form design good practice policy and procedures (GPPP)*.

P&P 3: Design the flow of data forms to entry to be:
• Continuous over the course of the trial
• Amenable to flow as soon as closed
• Independent of the flow of other forms
• The shortest, most direct, route to entry
Comment
  The requirements are necessary for ensuring timely data entry.

P&P 4: Design the data entry system to allow for all information contained on a form to be keyed in a single keying session; avoid systems in which forms are shunted to other persons to code and key selected items and systems where items containing narrative text (eg, responses to open-ended items) are flagged in lieu of being keyed.

Comment

   Flagging is done to simplify keying. Typically, it is done on the presumption that information contained in narrative responses will not have to be processed. If the presumption proves wrong (as is often the case), the cost and time required to identify and retrieve forms containing such responses will be far in excess of that involved if the information had been keyed straight away.

P&P 5: Opt in favor of double data entry whenever possible.

Comment

   The general standard is for double data entry.

P&P 6: When deciding on the type of double data entry, consider the cost and logistics of the different types (independent double data entry, quasi-independent double data entry, and dependent double data entry, as defined above).

Comment

   Generally, the simplest approach is to key and rekey in the same session (dependent double data entry). Independent double data entry is the "gold standard" but, of the three systems, is most difficult to manage.

P&P 7: Design and operate the data entry system to be WYSIWYE (what you see is what you enter) system.

Comment

   Key what is recorded, even if known to be wrong (see P&P 11).

P&P 8: Design the data entry system to keep track of keying sessions, keying strokes per unit of time, and keying error rates by keyer.

P&P 9: Establish and maintain an electronic inventory of forms.

Comment

   The forms inventory is useful in monitoring clinic performance. The inventory can be created from the database of keyed forms in systems where forms are keyed at the generation site. Otherwise it should be created by logging forms on receipt at the data entry site.

P&P 10: Use name code and Id number, or some other similar two variable system, for admission of records to the database.

Comment

   Linkage errors are serious and difficult to detect. Hence, the need for some reasonable fail-safe system to protect against them. Under such systems, a form is not added to the database if the linkage variables do not match.

Memo re Data entry good practice policies and procedures
(Monday 7:04am) 12 June 2000

P&P 11: Design the data entry system so as to be able to produce an audit trail of changes.
  Comment
   Changes to the study database should not be made without authorization and documentation. The documentation should be sufficient to allow one to reconstruct the sequence of events leading to the change, starting with the original value. Documents should indicate when and how the value came to be questioned (eg, the result of a routine edit done by the CC), how and when the query was relayed to clinic personnel (eg, by a dated, computer generated, edit query mailed or faxed to the clinic), the new value, the basis for the change, the date the change was authorized by clinic personnel, and the name or initials of the person in the clinic authorizing the change, and when the change was made to the data system.

P&P 12: Create the data system to include a field for version number and revision date of data collection forms.

P&P 13: Maintain an inventory of form versions and of procedures and programs for mapping the different versions of study forms to the study database.

\GPPP\DE.WPD