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

(Tuesday 9:21am) 26 December 2000

Memorandum

To: Center for Clinical Trials faculty and staff

Fr: Curt Meinert

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

Definitions

censor, **censored**, **censoring**, **censors** *v* - Broadly, to delete, suppress, or eliminate. rt: **informative censoring**, **uninformative censoring**, **mask** ***Usage note***: In the context of **trials**, censoring occurs when the **observation** of interest cannot be made or is not counted in an **analysis** because of some intervening condition or **event**, or in an effort to preserve **treatment masking** by withholding information related to an observation. An example of the latter kind of censoring is when certain laboratory determinations made on **study participants** and considered likely to show treatment-related effects are withheld from clinic personnel. Most censoring is of the former kind and arises from the fact that **enrollment** continues over a period of time and therefore persons at any point in time during the trial are seen for differing periods of time, depending on when enrolled. For example, suppose observation up to 30 Jan 1996 for an **interim analysis**, one person enrolled on 30 Nov 1995 (P_1), and one person enrolled on 30 Dec 1995 (P_2). P_1 contributes 60 person-days and P_2 contributes 30 person-days of observation. Observation of P_2 beyond day 30 is censored because of when enrollment occurred. A second form of censoring occurs because of **missed visits** or **dropouts**. For example, suppose it is not possible to observe P_1 beyond day 45 because the person refused further observation. Observation is censored at day 45 for variables requiring a **compliant person** for observation. Both forms of censoring arise from inability or failure to observe. Another form arises from eliminating observations made after occurrence of some event or condition. For example, suppose an analysis involving the **comparison** of **treatment groups** for an **event** (eg, the first occurrence of systolic blood pressure above a specified level) while on **assigned treatment**. Suppose that P_2 was taken off assigned treatment on day 15 and that the event of interest occurred on day 20. Observation, for purposes of the analysis, would be censored at day 15 and the event, though observed, would not be counted. Censoring not related to the variable of interest is referred to as **uninformative censoring** (eg, the censoring in the first example). Censoring related to the measure or event of interest is referred to as **informative censoring** (eg, censoring in the second example if **missed visits** are treatment-related).

clean data *n* - **Data** that have been subjected to **edit queries** and **editing** to reduce or resolve noted deficiencies due to missing information, **errors**, or inconsistencies. ant: **dirty data**

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

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 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.

dirty data *n* - 1. **Data** containing **errors** and deficiencies. 2. unedited data 3. Data with outstanding **edit queries** awaiting resolution. ant: **clean data**

final data analysis *n* - 1. **Data analysis** carried out at the end or last stage of a **study**, eg, analysis performed during the **termination stage** of a **trial**. 2. Data analysis performed in relation to the final version of a **manuscript** or **report**.

final dataset *n* - The **dataset** complied on completion of a **study** for use in **final data analysis** and for archiving.

outlier *n* - 1. Any value, reading, or **measurement** outside specified **limits**, especially one that is so far removed from other values that the appropriateness of its use in **analyses** is in question. 2. An extreme value most likely due to **error**.

Winsorization *n* - A procedure for reducing the influence of extreme values of a **continuous variable** on distribution-dependent **statistics**, such as **means** and **variances**; performed by ordering the observed values and then establishing **limits** defining the **right** and **left tails** of the **distribution** (eg, limits defined by the upper value of the first **decile** and the lower value of the

tenth decile of the distribution) below which values are considered extreme. All values to the left of the lower limit are assigned the value of the lower limit or the nearest **observed value** to the right of the limit; all values to the right of the upper limit are assigned the value of the upper limit or the nearest observed value to the left of the limit. Named for Charles P Winsor (1895 - 1951). See Hoaglin et al [1983] or Huber [1981] for details.

P&P 1: Establish procedures to ensure timely and continuous flow of data to the processing site.
Comment
Necessary for interim analyses and for ongoing performance monitoring.

P&P 2: Establish procedures for identifying dirty data (see defn above).
Comment
Having that means is important when creating and using analysis datasets; absent the means there is no way to restrict analyses to clean data (see defn above).

P&P 3: Establish policy for handling dirty data when creating analysis datasets; decide whether such data are to excluded from datasets when; if included decide whether such data are to be flagged (so they can be excluded from specified analyses).

P&P 4: Establish procedures for detecting and trimming outliers; perform trimming when creating frozen datasets.
Comment
The purpose of trimming is to reduce the influence of extreme values on means and variances and other distribution dependent statistics.

P&P 5: Establish procedures for data freezes; issues to be addressed include the following:
- Type and number of special edits to be performed before a freeze
- Data to be excluded from the frozen dataset
- Date of freeze relative to when the analysis or report is required
- Place of repose and storage of the frozen dataset

P&P 6: When preparing to produce datasets for manuscripts, allow sufficient lead time prior to the planned freeze date to allow for keying, harvesting, and editing data to be included in such datasets.

P&P 7: Document and store datasets supporting manuscripts; retain for a minimum of 3 years following publication of manuscripts.

P&P 8: In selecting variables for calculation of adjusted treatment comparisons, limit choice to variables observed at or prior to treatment assignment, limit number so as to have at least 20 treatment assignments per variable selected.
