Errors in individual measurements. If individual measurements are confirmed to be in error, make the necessary change and document it with your name and the date of the change. Verify all corrections from source information.

Systematic errors. Systematic errors are errors that occur consistently for a specific item. They often result from a misunderstanding or ambiguity in the protocol. If these errors are detected, try to ascertain whether the correct values can be deduced from the data or from source documents. If so, make the necessary calculations. Report the types of errors detected and the methods used to obtain comparable data.

Violation of rules. An example of this type of error would be if children older than a specified age were entered in the database. Always document changes.

Checklist

__ Check the source.
__ Develop a list of possible flaws.
__ Carry out range and consistency checks.

If you cannot detect any errors or discrepancies in the data, then you probably aren't looking hard enough.

Other brochures in this series

- Database ownership (1 of 3).
- Avoiding pitfalls that result in bad data (2 of 3).

Related brochures

- How Quantitative Health Sciences can satisfy your research needs.
- Sound principles for simple statistics.
- Working with spreadsheets.

QHS Section

Pippa M. Simpson, PhD
Director

Raymond G. Hoffmann, PhD
Associate Director

Shun-Hwa Li, PhD
Senior Biostatistician

Ke Yan, PhD
Senior Biostatistician

Mahua Dasgupta, MS
Biostatistician

Melodee Nugent, MA
Biostatistician

Chris Cronk, ScD
Senior Epidemiologist

JoAnn Gray-Murray, PHD
Qualitative Researcher

Database Support

Kathy Divine, MS
Database Administrator

Haydee Zimmerman, BA
Database Analyst II

Kim Gajewski, BA
Database Analyst II

Robert Thielke, PhD
Manager IS II

Children's Research Institute
A member of Children's Hospital and Health System.

Medical College of Wisconsin

Phone: (414) 955-7675 • Fax: (414) 955-6331
www.chw.org/qhs

Children's Research Institute
A member of Children's Hospital and Health System.

QHS
Quantitative
HEALTH SCIENCES

A good database doesn't mean good data

Guidelines for detecting bad data

Brochure 3 of 3

Quantitative Health Sciences was established to provide help in the design and analysis of research studies.

Detecting bad data

Too frequently, after data is collected there is little attempt to ascertain the reliability and validity of the data.

The integrity of a data set is a function of all aspects of data collection, entry and analysis. Any plan to ensure reliable and valid data should focus on preventing questionable practices, not only on ensuring data is entirely free of errors. The plan should use reasonable, cost-effective procedures to guarantee the validity of the primary results. There is no need to spend excessive funds to detect rare errors.

Variation is the spice of data. *Excessive variation, due to non-sampling error, often is the poison in data.*

Collecting and recording data

Data collection

- Who is responsible for collecting this data?
- During what period was it collected?
- Was there training for data extraction?

Data entry

- Was there training for data entry?
- When was the data entered into the database?
- Who entered the data?
- Was it double entered?
- Were there data entry checks or just editing for out-of-range variables?
- What was the philosophy about missing data?

Data flow

Documents should be available describing how data was collected. Look for a protocol, procedures manual or a coding book. If none of these exist, it is a strong sign for potential inconsistencies in the data. For example, weight often is miscoded as pounds when in fact the measurements are metric.

Data source

- Find out about the source of your data.
- Was it prospectively or retrospectively collected?
- Were values observed or measured? For example, was the figure for the weight obtained from a machine, by asking the patient or did the physician or nurse guess?
- Were definitions of terms unequivocally followed? For example, were the study guidelines followed when a patient was classified as severely ill?

Determining possible flaws in the data

Identify unusual patterns, lack of variability or unusual relationships in the data. Acceptable ranges of values:

- Are there extreme changes in values within subjects? Comparisons of values across time for a participant may reveal extreme values or a disturbing lack of variation.
- If it is feasible, return to the original source to detect possible errors. If a long series of data-processing steps occur between the source document and your database, check the values against documents at each level for at least a sample of cases.
- Measures of variability, central tendency and relationships of several variables may vary for different groups of observers, different places or even different demographic groups. If these measures are expected to be similar for each group, the existence of one group with significantly different measures might suggest a problem.

Missing data

- It usually is not feasible to fill the missing gaps, but the patterns of missing data often will fill in information.
- If there only is a small percent that do not have missing values for a variable, that variable probably should be discarded.
- If some group has more missing data than another the reason should be ascertained.
- If a particular value is unexpected or rarely seen it may indicate a problem.

Remember: Unusual statistical characteristics in the data do not necessarily imply a problem, but warrant further investigation. Many clinical observations are subjective and may be subject to inter- and intra-observer variability. Extreme changes in some measured values may be legitimate and even expected under the study conditions.

Possible solutions

Inform involved personnel. Document all changes. If there are a significant number of problems, you may consider developing a new database.