Data Cleaning, Linkage, and Data Analysis in Disease Registration

Ali Rafei
Program coordinator,
National Program for Disease Registries and Health outcomes
E-mail: a-rafei@alumnus.tums.ac.ir
Website: http://rafeistat.ir
Data cleaning
Sources of Error in Data

• Data entry errors

• Measurement errors

• Distillation errors

• Data integration errors
Data cleaning

- **Definition:** Detecting, Diagnosing, and Editing Data Abnormalities
- **Errors** occur in spite of careful study design, conduct, and implementation of error-prevention strategies
- Data cleaning intends to **identify** and **correct** these errors or at least to **minimize** their impact on study results
• Data handling, although having an equal potential to affect the quality of study results, has received proportionally less attention.

• Armitage almost apologized for inserting a short chapter on data editing in their standard textbook on statistics in medical research.

• However, data cleaning can never be a cure for poor study design or study conduct.
History

• **Statistical societies** recommend that description of data cleaning be a **standard part** of reporting statistical methods.

• In practice, it is **rare** to find any statements about **data-cleaning methods** or **error rates** in medical publications.

• Although **certain** aspects of data cleaning such as **statistical outlier detection** and **handling of missing data** have received separate attention, the data-cleaning process, **as a whole**, with all its **conceptual, organizational, logistical, managerial, and statistical-epidemiological** aspects, has not been described or studied comprehensively.
Terms Related to Data Cleaning

- **Data cleaning**: Process of detecting, diagnosing, and editing faulty data.

- **Data editing**: Changing the value of data shown to be incorrect.

- **Data flow**: Passage of recorded information through successive information carriers.

- **Inlier**: Data value falling within the expected range.

- **Outlier**: Data value falling outside the expected range.

- **Robust estimation**: Estimation of statistical parameters, using methods that are less sensitive to the effect of outliers than more conventional methods.
complete process of quality assurance

• error prevention
• data monitoring
• data cleaning
• documentation
Three-stage process of Data Cleaning

• repeated cycles of screening
• Diagnosing
• editing of suspected data abnormalities
A data-cleaning framework

Screening
- Lack/Excess of data
- Outliers/Inconsistencies
- Strange patterns
- Suspect analysis results

Diagnosis
- Errors, missing data
- True extreme
- True normal
- No diagnosis, still suspect

Editing
- Correction
- Deletion
- Leave unchanged

Design
Collect Enter
Transform Extract Transfer
Explore Analyze

System feedback
Study process

Data cleaning
Three-stage process of Data Cleaning

It is not always immediately clear whether a data point is erroneous.

Missing values:
- interruptions of the data flow
- unavailability of the target information

Predefined rules for dealing with errors and true missing and extreme values are part of good practice.
Three-stage process of Data Cleaning

The diagnostic and treatment phases of data cleaning require insight into the sources and types of errors at all stages of the study, during as well as after measurement.

<table>
<thead>
<tr>
<th>Data Stage</th>
<th>Sources of Problems: Lack or Excess of Data</th>
<th>Sources of Problems: Outliers and Inconsistencies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Questionnaire</td>
<td>Form missing</td>
<td>Correct value filled out in wrong box</td>
</tr>
<tr>
<td></td>
<td>Form double, collected repeatedly</td>
<td>Not readable</td>
</tr>
<tr>
<td></td>
<td>Answering box or options list left blank</td>
<td>Writing error</td>
</tr>
<tr>
<td></td>
<td>More than one option selected when not allowed</td>
<td>Answer given is out of expected (conditional) range</td>
</tr>
<tr>
<td>Database</td>
<td>Lack or excess of data carried over from questionnaire</td>
<td>Outliers and inconsistencies carried over from questionnaire</td>
</tr>
<tr>
<td></td>
<td>Form or field not entered</td>
<td>Value incorrectly entered</td>
</tr>
<tr>
<td></td>
<td>Data erroneously entered twice</td>
<td>Value incorrectly changed during previous data cleaning</td>
</tr>
<tr>
<td></td>
<td>Value entered in wrong field</td>
<td>Transformation (programming) error</td>
</tr>
<tr>
<td></td>
<td>Inadvertent deletions and duplications during database handling</td>
<td></td>
</tr>
<tr>
<td>Analysis dataset</td>
<td>Lack or excess of data carried over from database</td>
<td>Outliers and inconsistencies carried over from database</td>
</tr>
<tr>
<td></td>
<td>Data extraction or transfer error</td>
<td>Data extraction or transfer error</td>
</tr>
<tr>
<td></td>
<td>Deletions or duplications by analyst</td>
<td>Sorting errors (spreadsheets)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Data-cleaning errors</td>
</tr>
</tbody>
</table>

DOI: 10.1371/journal.pmed.0020267.t0001
Errors that need to be cleaned

• missing sex
• sex misspecification
• birth date or examination date errors
• duplications or merging of records
• and biologically impossible results

Example in nutrition: weight-for-age
Screening phase

four basic types of oddities:

• **lack** or **excess** of data
• **outliers**, including **inconsistencies**; **strange** patterns in (joint) distributions; and unexpected analysis results
• Other types of inferences and abstractions
Screening phase

identifying suspect data:

• normal ranges
• distribution shapes
• strength of relationships
Screening phase

erroneous inliers:

• data points generated by error but falling within the expected range
• Erroneous inliers will often escape detection

are discovered to be suspect if viewed in relation to other variables, using:

• scatter plots
• regression analysis,
• consistency checks
Screening methods

• Checking of questionnaires using fixed algorithms
• Validated data entry and double data entry
• Browsing of data tables after sorting
• Printouts of variables not passing range checks and of records not passing consistency checks
• Graphical exploration of distributions: box plots, histograms, and scatter plots
• Plots of repeated measurements on the same individual, e.g., growth curves
• Frequency distributions and cross-tabulations.
• Summary statistics
• Statistical outlier detection
Diagnostic Phase

the purpose is to **clarify** the true **nature** of the **worrisome** data points, **patterns**, and **statistics**.

Possible diagnoses for each data point:

- erroneous
- true extreme
- true normal (i.e. the prior expectation was incorrect)
- idiopathic (i.e. no explanation found, but still suspect)
Diagnostic Phase

Areas within the Range of a **Continuous** Variable Defined by **Hard** and **Soft** Cutoffs for Error Screening and Diagnosis, with Recommended Diagnostic Steps for Data Points Falling in Each Area
Graphical view of outliers

Multivariate Outlier Example

Visit Trends, by weekday and Hour

Hotelling T* Chart for Individuals
Diagnostic procedures

go to previous stages of the data flow to see whether a value is consistently the same

look for information that could confirm the true extreme status of an outlying data point

collect additional information, e.g., question the interviewer/measurer about what may have happened
Treatment phase

What to do?

- The options are limited to:
  - correcting
  - Deleting
  - leaving unchanged
Approaches to Improving Data Quality

What to do?

- Data entry interface design
- Organizational management
- Automated data auditing and cleaning
- Exploratory data analysis and cleaning
Documentation and Reporting

Good practice guidelines for data management require **transparency** and **proper documentation** of all procedures.

Data cleaning, as an **essential** aspect of **quality assurance** and a determinant of study validity, should not be an exception.

We suggest **including** a data-cleaning plan in **study protocols**.
Data linkage
Data linkage

• The process of linking and aggregating records from one or more data sources representing the same entity, like cancer registry data and death records.

• Also called data matching, data integration, data scrubbing, ETL (extraction, transformation and loading), object identification, merge-purge, etc.

• Challenging if no unique entity identifiers available e.g., which of these records represent the same person?

<table>
<thead>
<tr>
<th>Name</th>
<th>Address</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dr Smith, Peter</td>
<td>42 Miller Street 2602 O’Connor</td>
</tr>
<tr>
<td>Pete Smith</td>
<td>42 Miller St 2600 Canberra A.C.T.</td>
</tr>
<tr>
<td>P. Smithers</td>
<td>24 Mill Street 2600 Canberra ACT</td>
</tr>
</tbody>
</table>
Why do linkage?

- Augments available information for major diseases, risk factors, and health service utilization
  - Links exposures to outcomes
  - Provides longitudinal component to survey data

- Reduces cost burden
  - Re-contacting survey respondents for follow-up information can be expensive

- Increases accuracy and detail of data collected
Why do linkage?

Applications of data linkage:

- Remove duplicates in a data set (internal linkage)
- Merge new records into a larger master data set
- Compile data for longitudinal (over time) studies
- Geocode matching (with reference address data)

Widespread use of data linkage:

- Immigration, taxation, social security, census
- Fraud, crime and terrorism intelligence
- Business mailing lists, exchange of customer data
- Social, health and biomedical research
• **Techniques for record linkage**
  - Deterministic approach (unique ID exists)

  - Probabilistic approach

• **Probabilistic approach is built on five key components:**
  1. Define features that describe similarity between records.
  2. Place feature vectors into three classes: matches ($M$), non-matches ($U$), and possible matches ($P$).
  3. Perform record-pair classification by calculating the ratio $(P (Y | M)) / (P (Y | U))$ for each pair, where $Y$ is a feature vector for the pair and $P (Y | M)$ and $P (Y | U)$ are the probabilities of observing that feature vector for a matched and non-matched pair.
  4. Where no duplicate and/or non-duplicate record pairs are available, estimate conditional probabilities by using observed frequencies in the records to be linked.
  5. “Blocking,” or partitioning the databases based on some variable in both databases, improves efficiency.
Procedural issues in data linkage

Neither “data” nor “link” can be defined unambiguously, and the relationship between datasets can vary.

Linking vertically partitioned datasets carries little risk of re-identification, because in most cases there is no more information about a record on the combined dataset than was present in the individual datasets.

For horizontally partitioned datasets, it is necessary to link individual subjects’ records that are contained in two or more datasets. This process is risky because the combined dataset contains more information about each subject than either of the components.
Procedural issues in data linkage

Many linkage techniques depend on the presence of attributes in both databases that are unique to individuals but do not lead to re-identification.

Linkage can reduce data quality.

No matter how linkage is performed, other issues should be addressed:
- comparable attributes should be expressed in the same units of measure
- conflicting values of attributes for each individual common to both databases should be reconciled
- managing records that appear in only one database (most commonly they are dropped)
- consider effect of linkage on data quality

There are unremovable risks from data linkage. Strong consideration should be given to forms of data protection such as licensing and restricted access.
Considerations in data linkage

- Data linkage most often requires powerful computers and hardware
- Data linkage requires specific knowledge, skills and techniques
- Data linkage follows special rules to protect privacy, security and confidentiality of individual’s information
Examples of data linkage

- Hormone replacement therapy increases the risk of breast cancer
  - *Int J Cancer* 1999;81:339-44
- **Registers used:**
  - Prescription; cancer; death

- Low carbohydrate, high protein (Atkin's) diet increases risk of cardiovascular disease.
  - *BMJ* 2012; 344:e4026.
- **Registers used:**
  - Analytic cohort; death; migration; in-patient
Examples of data linkage

• Radical prostatectomy reduces mortality in patients with early prostate cancer
  • J Natl Cancer Institute 2008;100:1144-54
  • N Eng J Med 2011;364:1708-17

• Registers used:
  • Randomized trials; death

• Snus increases the risk of pancreatic cancer
  • Lancet 2007;369:2015-20

• Registers used:
  • Construction workers cohort; cancer; death
Data cleaning and linkage software

• Google Refine

• Talend Open Studio

• FRIL: Fine-Grained Records Integration and Linkage Tool

• LinkPlus
Data Entry software

- EPIData
- Microsoft Access
- Entry Point i4
- Google form
- Questys Capture
Data Analysis
Data analysis

Analysis and interpretation of registry data begin with a series of core questions:

• Study purpose
• Patient population
• Data quality
• Data completeness
• Data analysis
Hypotheses and purpose

**Descriptive:**

- describing the typical clinical features of individuals with a disease
- variations in phenotype
- the clinical progression of the disease over time

**Association:**

Outcome vs exposures
Patient population

Representativeness:

describe the demographics and other key descriptors of the registry study population and to contrast its composition with patients with similar characteristics who are identified from an external database

- **Example:** Diabetes drug in adults
- **Example:** isolated population

If the action of the drug (or its delivery) does not vary geographically

**Intended population:** The main difference is that the intended population may be specified by a sampling scheme, which often tries to strike a balance among representativeness, convenience, and budget
Data quality issues

Collection of All Important Covariates
• provider performance feedback
• addressing a specific clinical research question

Data Completeness

Missing Data:
• MCAR
• MAR
• MNAR
Missing data

- **Complete case strategy** (used by many software)
  A simple deletion of all incomplete observations, however, is **not appropriate**
Example in follow-up studies: Diabetes with HA1c

**Imputation** techniques for MNAR:
- Unconditional or conditional mean
- Multiple hot-deck
- Expectation maximization
- Multiple imputation
Data analysis

Descriptive statistics:
Measures, Tables, and Charts

Variables:
Continuous (mean, histogram, frequency table)
Categorical (Incidence, prevalence, median, mode, bar-chart, contingency table)
Association measures

• Attributable risk (AR)
• Relative risk (RR)
• Odds ratio (OR)
• Excess risk (ER)

Role of:
• Confounders
• Effect-modifiers
Confounding

- Stratified analysis
- Multivariable analysis
- Sensitivity analysis
- Simple or quantitative bias analysis

New methods:
high-dimensional propensity score (hd-PS) for adjustment using administrative data
Grouping or Clustering

• Analysis of Variance

• Hierarchical or Multilevel modeling
Other issues

Heterogeneity of treatment effect
• Stratification on the propensity score

Economic analysis
• cost-effectiveness analyses
• cost-utility analyses
The flow of participants into an analysis

Potential participants assessed for eligibility ($n = ...$)

- Excluded ($n = ...$)
  - Ineligible $n =$
  - Reasons... $n =$

Eligible ($n = ...$)

- Did not consent ($n = ...$)
  - Refused $n =$
  - Other reasons... $n =$

Consent to participate ($n = ...$)

Only required if numbers consenting are not the same as the numbers at baseline

Numbers participating at baseline data collection ($n = ...$)

- Losses after consent ($n = ...$)
  - Reasons... $n =$

Numbers participating at $n$th waves of data collection ($n = ...$)

Only required if $>1$ follow-up

Losses after follow-up ($n = ...$)
  - Reasons... $n =$

Numbers participating at final wave of data collection ($n = ...$)
Developing a Statistical Analysis Plan

Need for a Statistical Analysis Plan
analytical principles and statistical techniques
Master SAP vs subsequent, supplemental SAP

Preliminary Descriptive Analysis To Assist SAP Development
Comprehensive Statistical Software
Other Packages:

- CDC Epi Info: Field epidemiology
- JMP: Visual analysis and statistics
- Systat: General statistical package
- TreeAgePro: Decision science
- SUDAAN: Survey data
- WinBUGS: Markov chain Monte Carlo
- Python: General-purpose programming language
- Matlab, Mathematica: Mathematics
<table>
<thead>
<tr>
<th>Product</th>
<th>Developer</th>
<th>Latest version</th>
<th>Cost (USD)</th>
<th>Open source</th>
<th>Software license</th>
<th>Interface</th>
<th>Written in</th>
<th>Scripting languages</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAS</td>
<td>SAS Institute</td>
<td>July 2013</td>
<td>Academics and colleges students: free Commercial: ~$6000 per seat (PC version) / ~$28K per seat</td>
<td>No</td>
<td>Proprietary</td>
<td>CLI/GUI</td>
<td></td>
<td>SAS language</td>
</tr>
<tr>
<td>SPSS</td>
<td>IBM</td>
<td>August 2013</td>
<td>$1599</td>
<td>No</td>
<td>Proprietary</td>
<td>CLI/GUI</td>
<td>JAVA</td>
<td>R, Python, SaxBasic</td>
</tr>
<tr>
<td>Stata</td>
<td>StataCorp</td>
<td>June 24, 2013</td>
<td>academic starting at $595 industry starting at $1,245</td>
<td>No</td>
<td>Proprietary</td>
<td>CLI/GUI</td>
<td></td>
<td>C ado, mata</td>
</tr>
<tr>
<td>R</td>
<td>R Foundation</td>
<td>March 6, 2014</td>
<td>Free</td>
<td>Yes</td>
<td>GNU GPL</td>
<td>CLI/GUI</td>
<td>C</td>
<td>Python, Perl</td>
</tr>
<tr>
<td>S-Plus</td>
<td>Insightful Inc.</td>
<td>2010</td>
<td>$2399/year</td>
<td>No</td>
<td>Proprietary</td>
<td>CLI</td>
<td></td>
<td>S language</td>
</tr>
<tr>
<td>Minitab</td>
<td>Minitab Inc.</td>
<td>February 18, 2010</td>
<td>$895–$1395</td>
<td>No</td>
<td>Proprietary</td>
<td>CLI/GUI</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Statistica</td>
<td>Dell/StatSoft</td>
<td>April 2013</td>
<td>$1100/year</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Epi Info</td>
<td>CDC</td>
<td>January 26, 2011</td>
<td>Free</td>
<td>No</td>
<td>Proprietary</td>
<td>CLI/GUI</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Operating system support

<table>
<thead>
<tr>
<th>Product</th>
<th>Windows</th>
<th>Mac OS</th>
<th>Linux</th>
<th>BSD</th>
<th>Unix</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAS</td>
<td>Yes</td>
<td>Terminated</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>SPSS</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Stata</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>R</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>SPlus</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Minitab</td>
<td>Yes</td>
<td>Terminated</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Epi Info</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>

### Support for various ANOVA methods

<table>
<thead>
<tr>
<th>Product</th>
<th>One-Way</th>
<th>Two-Way</th>
<th>MANOVA</th>
<th>GLM</th>
<th>Post-hoc Tests</th>
<th>Latin Squares Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAS</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>SPSS</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Stata</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>R</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>SPlus</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Minitab</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Epi Info</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>
### Support for various regression methods

<table>
<thead>
<tr>
<th>Product</th>
<th>OLS</th>
<th>WLS</th>
<th>2SLS</th>
<th>NLLS</th>
<th>Logistic</th>
<th>GLM</th>
<th>LAD</th>
<th>Stepwise</th>
<th>Quantile regression</th>
<th>Probit</th>
<th>Poisson</th>
<th>MLR</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAS</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>SPSS</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stata</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>R</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>SPlus</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Minitab</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Epi Info</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Support for various statistical charts and diagrams

<table>
<thead>
<tr>
<th>Chart</th>
<th>Bar chart</th>
<th>Box plot</th>
<th>Correlogram</th>
<th>Histogram</th>
<th>Line chart</th>
<th>Scatterplot</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAS</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>SPSS</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Stata</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>R</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>SPlus</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Minitab</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Epi Info</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>
### Other capabilities

<table>
<thead>
<tr>
<th>Product</th>
<th>S/W type</th>
<th>Descriptive</th>
<th>Nonparametric</th>
<th>Quality Control</th>
<th>Survival Analysis</th>
<th>Data Processing</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAS</td>
<td>S</td>
<td></td>
<td></td>
<td>+</td>
<td></td>
<td>+</td>
</tr>
<tr>
<td>SPSS</td>
<td>S</td>
<td></td>
<td></td>
<td>+</td>
<td></td>
<td>+</td>
</tr>
<tr>
<td>Stata</td>
<td>S</td>
<td></td>
<td></td>
<td>+</td>
<td></td>
<td>+</td>
</tr>
<tr>
<td>R</td>
<td>S</td>
<td></td>
<td></td>
<td>+</td>
<td></td>
<td>+</td>
</tr>
<tr>
<td>SPlus</td>
<td>St</td>
<td></td>
<td></td>
<td>+</td>
<td></td>
<td>+</td>
</tr>
<tr>
<td>Minitab</td>
<td>S</td>
<td></td>
<td></td>
<td>+</td>
<td></td>
<td>+</td>
</tr>
<tr>
<td>Epi Info</td>
<td>S</td>
<td></td>
<td></td>
<td>+</td>
<td></td>
<td>-</td>
</tr>
</tbody>
</table>

CTA = Contingency Tables Analysis  
S = Standalone executive; St = Standalone executive  
BDP = Base Data Processing, f.ex. Sorting  
Ext. = Extended (data sampling, transformation)