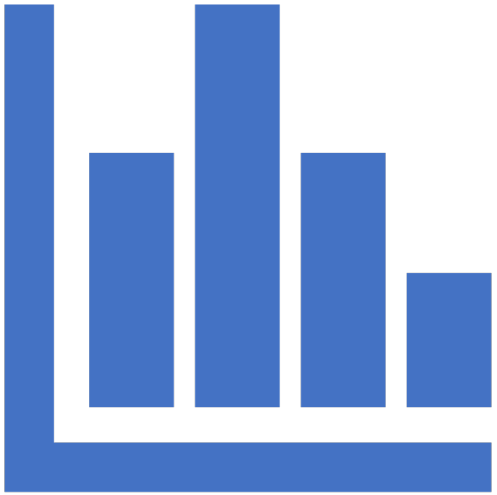


Data Cleaning

Table 1. Issues to Be Considered during Data Collection, Management, and Analysis of a Questionnaire Study

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

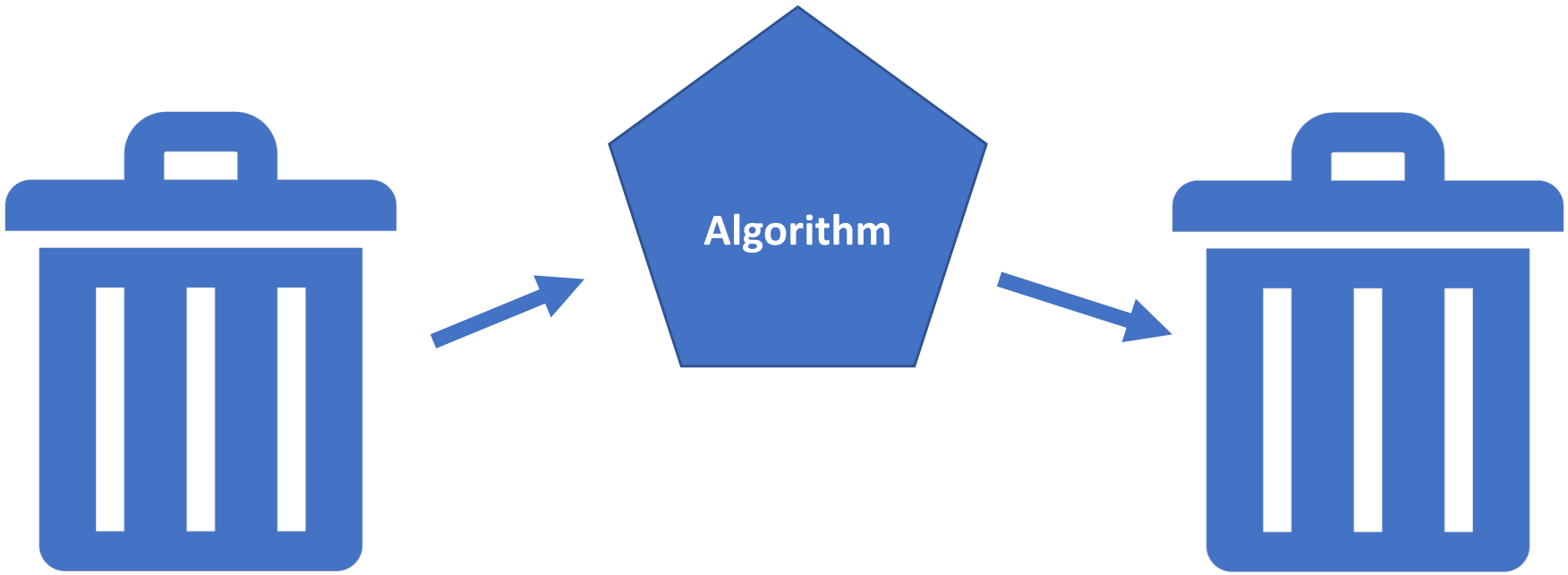
DOI: 10.1371/journal.pmed.0020267.t001

Van den Broeck J, Argeseanu Cunningham S, Eeckels R, Herbst K (2005) Data Cleaning: Detecting, Diagnosing, and Editing Data Abnormalities. PLOS Medicine 2(10): e267. https://doi.org/10.1371/journal.pmed.0020267
https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.0020267

Do not fake your data.
This is not that.

Always report cleaning.

Does *Big Data* eliminate the need for data cleaning?



Data challenges

- Non-standardized data
- Inconveniently structured data
- Duplicate data
- Missing values
- Data with multiple factors
- Incorrect values