Datenerfassung

Hintergrund

Um für einen einheitlichen Datenaufbau zu sorgen und die anschließende Auswertung von erfassten Daten zu erleichtern, ist eine vorherige Planung vorzunehmen.

Dieser Leitfaden soll dabei unterstützen, diese Planung möglichst ideal durchzuführen. Es wird dadurch ein (aufwändiges) Nachbearbeiten der Daten vermieden oder zumindest erleichtert.

Datenerfassungsprogramme

Zur primären Datenerfassung und/oder –export eignen sich:

- Tabellenkalkulationsprogramme (z.B. MS Excel, Open Office Calc, Google Sheets)
- Datenbanken (z.B. MS Access, MySQL)
- Statistikprogramme /-umgebungen (z.B. SPSS)

Daten

Grundregeln zum Aufnehmen von Daten umfassen:

- jeder Datensatz/Proband[1] erhält eine eindeutige Identifikationsnummer (ID)
- jede Information, die zu einem Datensatz/Probanden erfasst werden soll, benötigt mindestens eine eigene Variable
- es wird pro Variable nur eine Information erfasst (siehe Beispieltabelle die Variable *Größe*, es wird nur der nummerische Wert gespeichert – nicht die Messeinheit)
- Variablennamen sind einzigartig, möglichst aussagekräftig zu benennen und kurz zu halten (10-12 Zeichen); sie enthalten keine Leerzeichen und beginnen mit einem Buchstaben oder einer Ziffer. Als erlaubtes Sonderzeichen eignet sich u.a. ein Unterstich _
- werden die Daten nicht im Rahmen einer klinischen Studie erhoben, ist das Geburtsjahr des Probanden als Information ausreichend

In Bezug auf Tabellenkalkulationsprogramme, sollte generell vermieden werden Formatierungen (z.B. Fettdruck, farbig hinterlegte Zellen etc.) und zusätzlich berechnete Zellen (z.B. mit Alters-berechnungen oder Mittelwerten) zu nutzen.

Probandennamen sind aus Datenschutzgründen **nicht** zu erfassen bzw. getrennt von den zu analysierenden medizinischen Daten zu speichern.

Werte und Kodierung

Die Ausprägungen kategorialer Merkmale sind mit numerischen Werten zu kodieren. Beispielsweise biologisches Geschlecht nicht in Textform sondern in den Ziffern *0* (= männlich), *1* (= weiblich) und 2 (= divers). Dadurch wird das analysieren der vorhandenen Informationen ermöglicht.

Die Kodierung von Variablen und deren Werten muss dokumentiert werden und sollte, wie auch die Variablennamen, möglichst kurz und knapp sein (siehe Code-Plan Beispiel). Um fehlende Werte zu kodieren muss ein Wert außerhalb des natürlichen Wertebereichs der Variable verwendet werden. Für die meisten Variablen bietet sich -9 an. Datumsvariablen (Geburtsdatum, Studieneintritt, …) und

[1] Die gewählte männliche Form bezieht sich immer gleichermaßen auf weibliche und männliche Personen.

bedingt fehlende Werte (Proband ist beispielsweise nicht zur Untersuchung erschienen) sind davon ausgenommen, fehlende Werte werden in diesen Fällen nicht kodiert und bleiben fehlend.

Durch die Dokumentation der Kodierung (auch Data Dictionary oder Code-Plan genannt) sind die Daten später korrekt zu interpretieren. Hier werden auch die Messeinheiten von Variablen dokumentiert (siehe Beispiel). Mittels Syntax kann die Kodierung (Variablenlabel, Wertelabel und fehlende Werte) in SPSS übernommen werden.

Tabelle 1: Beispiel eines Code-Plans

Variablenname	patienten_id	geschlecht	gbdat	gewicht	gruppe	sysbp
Variablenlabel	Patienten Identifikation	Geschlecht	Geburtsdatum	Gewicht in kg	Behandlungs gruppe	Systolischer Blutdruck in mmHg
Kodierungen	<Wert>	0 = männlich	TT.MM.JJJJ	<Wert>	1 = Verum	<Wert>
	<Wert>	1 = weiblich	TT.MM.JJJJ	<Wert>	0 = Placebo	<Wert>

Berechnungen und Subgruppen

Für neu zu berechnende Variablen, sollte immer SPSS genutzt werden. Berechnungen sind mittels Syntax dokumentiert und somit reproduzierbar, wodurch die Analysen auch für Dritte nachvollziehbar bleiben und so außerdem den gesetzlichen Regularien gefolgt wird (z.B. GCP, GEP, ICH, …). Das betrifft beispielsweise den BMI, der aus Größe und Gewicht eines Probanden berechnet werden kann, die Altersberechnungen, sowie eine Variable zur Subgruppenbildung (z.B. Probanden die einen BMI über 35 haben und männlich sind).

Aufnahmeformate

Zur Planung der Datenerfassung und deren Formaten, ist die Fragestellung zu beachten, die mit den Daten beantwortet werden soll. Für Daten mit Messwiederholungen (pro Proband mehrere Beobachtungen zu einer Messgröße) gibt es zwei Varianten, die Datendatei aufzubauen.

Wide-Format:
Ein Proband pro Zeile, mit mehreren Beobachtungen einer Messgröße in eigenen Variablen. T steht für Zeitpunkt.

Tabelle 2: Beispiel Datenerfassung im Wide-Format

probanden_id	groesse	gewicht_t1	gewicht_t2	sysbp_t1	sysbp_t2
1	181	80	70	117	110
2	176	90	75	111	105

Long-Format:
Mehrere Zeilen pro Proband, mit einer Variablen für alle Beobachtungen der Messgröße.

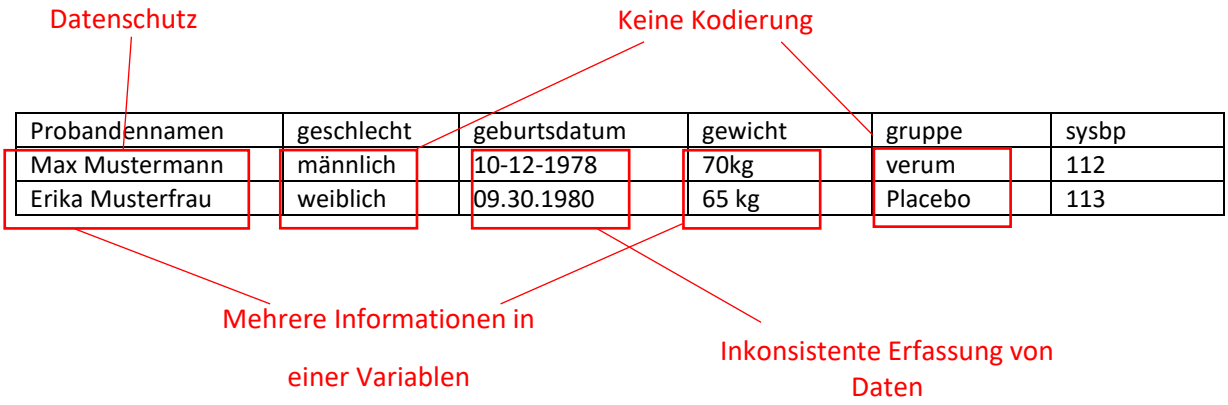
Tabelle 3: Beispiel Datenerfassung im Long-Format

probanden_id	Zeitpunkt	groesse	gewicht	sysbp
1	1	181	80	117
1	2	181	70	110
2	1	176	90	111
2	2	176	75	105

Ergeben sich keine Messwiederholungen in den Daten, wird grundsätzlich das Wide-Format verwendet. Also ein Proband pro Zeile. Bei Messwiederholungen ist das Long-Format zu bevorzugen.

Beispiele die zu vermeiden sind

Tabelle 4: Beispiel für fehlerhafte Datenerfassung



Probandennamen	geschlecht	geburtsdatum	gewicht	gruppe	sysbp
Max Mustermann	männlich	10-12-1978	70kg	verum	112
Erika Musterfrau	weiblich	09.30.1980	65 kg	Placebo	113

Richtig:

Tabelle 5: korrigierte Datenerfassung

patienten_id	Geschlecht	geburtsdatum	gewicht	gruppe	sysbp
1003	0	10.12.1992	70	1	112
1004	1	30.09.1980	65	0	113