Planung der Dateneingabe

Wie kodiere ich meine Variablen?

Hier muss ich bereits wissen, welche Informationen (Variablen) ich erfassen will.

Zum Beispiel:

- Alter
- Geschlecht
- Familienstand
- Verkehrsmittel zur Klinik
- Raucherstatus
- Blutgruppe
- Puls
- Blutdruck
- Immunglobulin

Dann stellen sich für jede Variable die folgenden Fragen:

Ist die Antwort eine Zahl?
Wie genau will ich diese Zahl erfassen? (z.B. Anzahl an Nachkommastellen, Rundung)

Ist die Antwort ein Text, der sich gut als Zahl kodieren lässt? (begrenzte Anzahl an Antwortkategorien, (fast) alle Antwortkategorien im Vorhinein bekannt?)
Brauche ich eine Antwortkategorie „Sonstiges“?
Will ich, dass bei Angabe von „Sonstiges“ dies im Freitext näher spezifiziert wird?
Ist genau eine Antwortkategorie möglich, oder können mehrere Antwortkategorien gleichzeitig zutreffen?

Ist die Antwort ein („unvorhersehbarer“) Freitext?

Wie kodiere ich fehlende Werte? Will ich verschiedene Arten von fehlenden Werten unterscheiden können (z.B. weiß nicht, trifft nicht zu, Antwort verweigert, Messung fehlt, Messung liegt noch nicht vor, Messung nicht durchführbar)

Wie viele Variablen brauche ich, um einen Inhalt sinnvoll zu erfassen?

Erstellung eines Codebuches

Das Codebuch enthält die Dokumentation zu meinen Daten.
(siehe Beispieldatei „Codebuch.xlsx“)

- Verwendete Sprache?
- Festlegung von ID-Variablen
- Anonyme (Haupt-) Daten
- Eigentlich interessierende Variablen
- Kodierung fehlender Werte

Verwendete Sprache?

- Deutsch oder Englisch?
- In welcher Sprache wird die (Doktor-) Arbeit verfasst?
- Wird die Arbeit publiziert? In welcher Sprache?
- In welcher Sprache ist die existierende Literatur zum Thema verfasst?

(Insbesondere relevant für Variablenlabel und Wertelabel)

Festlegung von ID-Variablen

Welche Identifikations-Variablen brauche ich?
Wie kann ich einen Patienten eindeutig identifizieren?

Vorschläge:

- Eine mehrstellige Zahl, z.B. eine dreistellige Zahl bei bis zu 999 Patienten (pro Studienzentrum)
- Bei mehreren Studienzentren: eine ID für das Studienzentrum / die Klinik / die Arztpraxis
- Der Erhebungszeitpunkt (bei Längsschnittstudien), z.B. 0, 1, 2, 3 oder Datum

Welche weiteren Variablen könnten hilfreich sein?

Weiterhin evtl. hilfreich, z.B. wenn man später „Fehler“ oder „Ungereimtheiten“ in den Daten sucht / findet (aber auch für die Analyse der Daten):

- Ort / Land (z.B. Übersetzungsfehler bei mehreren Sprachen)
- Datum der Erhebung
- ID des Interviewers (eventuelle „Eigenheiten“ des Interviewers)
- Datum der Dateneingabe
- ID desjenigen, der die Daten eingibt (Beispiel: einer hat sich die Kodierung falschrum gemerkt, einer arbeitet sehr schlampig)

Gute Praxis: Anonyme Daten

Wie stelle ich sicher, dass meine „Hauptdaten“ möglichst anonym sind, ich aber trotzdem noch nachschauen kann, wer wer ist?

In den **Hauptdaten** sollten nur **anonyme IDs** gespeichert werden, nicht die Namen oder Adressen der Patienten. (Deshalb wird auch die Erfassung des Geburtsdatums von der Ethik-Kommission teilweise nicht genehmigt.)

Die Zuordnung von IDs zu den Patienten erfolgt in einer **separaten Datei**.
(Nicht in einem neuen Tabellenblatt derselben Excel-Datei wir die Hauptdaten!)
(siehe Beispieldatei „Studienteilnehmer.xlsx“)

In den ersten Spalten sollten **alle benötigten ID-Variablen** stehen.
In den weiteren Spalten sollten alle Angaben erfasst werden, die zur **Identifizierung** und zur **Kontaktierung** des Patienten notwendig sind.

Beispiele:

- Name
- Geburtsdatum
- Versicherungsnummer
- Klinik-interne Patienten-ID
- Auftragsnummer / Fallnummer
- Station

- Adresse, Telefonnummer, Email-Adresse

- und andere Angaben (z.B. behandelnder Arzt, Hausarzt)

Festlegung der eigentlich interessierenden Variablen – im Codebuch

(siehe Beispieldatei „Codebuch.xlsx“)

- Variante 1: für SPSS benötigte Angaben – Ausprägungen untereinander
- Variante 2: für SPSS benötigte Angaben – Ausprägungen nebeneinander
- Variante 3: Minimalversion – nicht wirklich sinnvoll
- Variante 4: ausführliche Version mit Zusatzangaben

- **Variable / Inhalt**
  Diese Spalte ist optional. Sie bietet sich an, wenn ein Inhalt in mehreren Variablen gespeichert werden muss. Sie kann auch dazu verwendet werden, um eine „interne“ Bezeichnung hineinzuschreiben, also wie diese Variable im Sprachgebrauch (z.B. unter Kollegen) genannt wird.

- **Variablenname**
  eher kurz, aber trotzdem möglichst selbsterklärend;
  keine Leerzeichen, keine Sonderzeichen (Unterstriche ok), (möglichst) keine Umlaute;
  möglichst einheitlich deutsch oder englisch

- **Variablenlabel (deutsch)**
  ausführlichere Beschreibung der Variable, mit Angabe der Einheit
  ausführlicher, "hübscher" Text, z.B. direkt geeignet zur Beschriftung von Grafiken – deshalb gut überlegen!
  Leerzeichen und Umlaute ok, Sonderzeichen möglichst vermeiden
  möglichst keine unnötigen Leerzeichen (am Anfang, innerhalb des Textes, am Ende)

- **Variablenlabel englisch - optional**
  wie Variablenlabel (deutsch)

    - wenn alle Ergebnisse ausschließlich auf Deutsch oder Englisch präsentiert werden sollen, reicht eine der beiden Spalten

- **Zusatzinformationen / Vorwissen - optional**
  Zusatzinformationen bzw. Vorwissen zur Variable, z.B.
    - Begrenzungen des Wertebereichs von metrischen Variablen (z.B. Patienten erst ab einem gewissen Alter in die Studie eingeschlossen)
    - Normwerte bzw. Wertebereich „normaler“ Werte
    - ab welchem Wert handelt es sich vermutlich um Messfehler?
    - Informationen zur Interpretation der Werte, z.B. umso höher der Wert, umso besser die Gesundheit des Patienten
    - Information zur Berechnung der Variable, z.B. Summe aus drei anderen Variablen

- **Skalenniveau** - mit den Optionen:
    - Nominal - Nominalskala:
        - Ausprägungen sind nur Begriffe
        - keine Ordnung/Reihenfolge vorhanden
        - Beispiele: Geschlecht, Familienstand, Augenfarbe, Blutgruppe
    - Ordinal – Ordinalskala
        - Ausprägungen lassen sich in einer Rangfolge anordnen
        - es gibt eine Kleiner-Größer-Beziehung
        - die Abstände sind nicht interpretierbar
        - Beispiele: Schulnoten, Tumorgröße, Lebensqualität, medizinische Scores
    - Metrisch
        - Ausprägungen unterscheiden sich zahlenmäßig
        - Differenz kann berechnet werden
        - Beispiele: Alter, Körpergröße, Gewicht, Leukozytenzahl
    - Evtl. Zusatzangaben wie „Freitext“ (eigentlich nominal), „Datum“ (eigentlich metrisch), usw.

- **für metrische Variablen**
    - Anzahl Nachkommastellen
    - Plausibles bzw. theoretisch mögliches Minimum – falls vorhanden (zur Plausibilitätsprüfung) - optional
    - Plausibles bzw. theoretisch mögliches Maximum – falls vorhanden (zur Plausibilitätsprüfung) - optional

- **für kategoriale Variablen** (nominal und ordinal)
    - Antwortkategorien (Zuordnung von Zahl zu Bedeutung, z.B. 1=männlich, 2=weiblich)
    - falls Mehrfachantworten zulässig sind, eine Variable pro Antwortkategorie
    - Bei Freitext: evtl. Anzahl an Zeichen
    - Kodierung (für den Inhalt vergebene Kodierung (meist Zahl), auch direkte Eingabe von (Frei-)Text möglich)
    - Bedeutung (Bedeutung der eingegebenen Zahl)
    - Bedeutung englisch - optional

- **für fehlende Werte**
    - evtl. Kodierung für fehlende Werte (damit man weiß, dass man nicht nur aus Versehen nichts eingegeben hat)
    - evtl. unterschiedliche Kodierung für verschiedene Gründe für fehlende Werte
    - Kodierung (für den Inhalt vergebene Kodierung (meist Zahl), auch direkte Eingabe von (Frei-)Text möglich)
    - Bedeutung (Bedeutung der eingegebenen Zahl)
    - Bedeutung englisch - optional

Das Codebuch kann direkt in SPSS erstellt werden – empfehle aber die Erstellung des Codebuches in Excel (auch wenn man das dann doppelt eingeben muss)

Kodierung fehlender Werte – im Codebuch

(Muss im Codebuch mit dokumentiert werden!)

Oft verwendet: 9, 99, 999, 9999, je nach Anzahl der Stellen der Variable

- sinnvoll, falls jeweils nur genau diese Anzahl an Stellen im gewählten Datenformat gespeichert werden können
- z.B. beim Speichern der Angaben in einer reinen Textdatei, sinnvoll bei sehr großen Datensätzen
- Nachteil: Man muss für jede Variable einzeln die fehlenden Werte spezifizieren.

Alternative: z.B. 99999 für alle Variablen

- weniger Arbeit bei der Kennzeichnung / Berücksichtigung der fehlenden Werte im Datensatz
- Nachteil: größerer Speicherplatzbedarf, „intelligenteres“ Datenformat notwendig

- Wert sollte so groß sein, dass es bei der Analyse auffällt, wenn die fehlenden Werte nicht richtig berücksichtigt werden.
- Beispiel für ungeeignete Kodierung des fehlenden Wertes: 99 bei Alter ist ein theoretisch möglicher Wert (ähnliches gilt für Größe, Gewicht, sonstige Messungen, Blutparameter, Laborwerte,…)
- Besser übertrieben groß wählen!

- Alternativ: bei metrischen Variablen, die nur positive Werte annehmen können, fehlende Angaben mit negativen Werten kodieren, z.B. -1, -2, -3, usw.

Evtl. unterschiedliche Kodierungen für verschiedene Ursachen für die fehlenden Werte wählen.

- Bei Fragen:
  - Antwort verweigert / fehlt
  - Weiß nicht
  - Frage nicht anwendbar / Sachverhalt trifft auf den Patienten nicht zu

- Bei Messwerten:
  - Wert fehlt (kann nicht mehr nachträglich ermittelt werden)
  - Wert kann noch nachgetragen werden / Laborergebnisse stehen noch aus, …
  - Wert ist für den Patienten nicht relevant / nicht anwendbar
  - Wert darf für den Patienten nicht ermittelt werden (Messung nicht durchführbar)