ZENTRUM FÜR MEDIZINISCHE STATISTIK,
INFORMATIK UND INTELLIGENTE SYSTEME
MEDIZINISCHE UNIVERSITÄT WIEN
Institut für Klinische Biometrie

Anforderungen an Datensätze zur statistischen Auswertung

Sehr geehrte Kooperations-PartnerInnen,

Im Sinne einer *Good Scientific Practice* stellen wir folgende Anforderung an Datensätze, die uns zur statistischen Auswertung übergeben werden. Diese Anforderungen sind aus jahrelanger Erfahrung im Umgang mit Datensätzen entstanden. Wir ersuchen Sie, diese möglichst genau zu berücksichtigen. Sie erleichtern uns damit die Einlesbarkeit und Interpretation der Daten und beschleunigen somit unsere Zusammenarbeit. Vielen Dank!

Das Wichtigste zusammengefasst:

- Daten aus SPSS oder Datenbank (z.B. ClinCase, MS Access) bevorzugt
- Variablennamen eindeutig, kurz, ohne Umlaute, Leer- und Sonderzeichen
- ID-Variable, keine PatientInnen-Namen, -Adressen oder -Telefonnummern
- Numerische Variablen (Spalten) enthalten ausschließlich Zahlen
- Data Dictionary zur Erklärung und Differenzierung von Variablen
- Daten-Änderungen nach Abschluss des Data Cleanings vermeiden

Software

1. **ClinCase**

   Für prospektive Studien wird die Verwendung des an der MedUni Wien durch das ITSC zur Verfügung gestellten Produktes **ClinCase** empfohlen. ClinCase erlaubt die Erstellung von Eingabemasken, Abfangen von Fehleingaben, Auditing etc. und erhöht damit die Datenqualität. Kontakt: melanie.fraunschiel@meduniwien.ac.at
2. **Bevorzugung von SPSS** gegenüber MS Excel
   a. Die Eingabe der Daten in SPSS oder Datenbanksystemen (z.B. MS Access) wird klar bevorzugt, da diese Daten leichter statistisch verarbeitbar sind.
   b. Möglichkeit von **Werte-Labels** in SPSS, um numerischen Codes für Kategorien von qualitativen Variablen Bedeutungen zuzuordnen (z.B. Geschlecht als numerische Variable mit Codes 0 und 1 und Wertelabels 0=weiblich, 1=männlich)
   c. Möglichkeit von **Variablen-Labels**, um kurzen, prägnanten Variablennamen genauere Informationen zuzuordnen (z.B. Variablenname = „blutdruck0“, Variablen-Label =“syst. Blutdruck zu Baseline (mmHg)“)
   d. In Spalten für **numerische Variablen** in SPSS können nur Zahlen eingegeben werden, was Eingabefehler vermeidet und die Einlesbarkeit erleichtert.
   e. **Fehlende Werte** sind in SPSS eindeutig als solche erkennbar.

Variablen

3. **Variablennamen**
   a. Eindeutig, möglichst kurz und leicht interpretierbar
   b. Keine Umlaute, Leer- oder Sonderzeichen (Unterstrich ist erlaubt)
   c. Erstes Zeichen des Variablennamens ist ein Buchstabe (A-Z).
   d. Die Variablennamen selbst enthalten keine Codierung (z.B. „geschlecht“ statt „Geschlecht 0=männl, 1=weibl“).
   e. Möglichst keine physikalischen Einheiten im Variablennamen

Spitalgasse 23 · A-1090 Wien
Telefon +43 1 40400-66880 · Fax +43 1 40400-66870
biometrie@meduniwien.ac.at · http://www.meduniwien.ac.at/cemsiis/kb/
DVR-Nummer 2108638

4. **Wiederholte Messungen**
   a. „Langes Dateiformat“ (eine Zeile pro Messung) benötigt eine Variable/Spalte zur Identifikation der Messung (z.B. Visite mit Werten 1, 2, 3,…)
   b. „Breites Dateiformat“ (eine Zeile pro Pat.) benötigt eine möglichst fortlaufende Nummerierung der Spalten, z.B. Blutdruck0, Blutdruck1, Blutdruck2
5. **Mehrfachnennungen**
   a. Für Fragen mit möglichen Kombinationen von Antworten (z.B. 10 Diagnosen, Mehrfachdiagnosen treten auf).
   b. Für jede Antwort-Kategorie (z.B. Diagnose 1 = Diabetes) muss eine eigene Variable (z.B. „Diabetes“) definiert werden, die jeweils 0 (= Nein) oder 1 (= Ja) enthält.
   c. Eingaben wie „1 + 5“ (die als fehlende Werte interpretiert würden) werden somit vermieden.
6. **PatientInnen-Identifikation**
   a. Die erste Variable im Datensatz ist eine eindeutige **Pat.identifikation** (Nummer oder Code), ggf. in Übereinstimmung mit dem Code auf dem Datenerhebungsbogen.
   b. Pat.namen, Adressen, Telefonnummern etc. sind aus dem Datensatz zu entfernen
7. **Datenlayout in Excel** (falls SPSS oder Datenbank-Systeme nicht verfügbar)
   a. Alle Daten sollten möglichst in einem **einzigen Tabellenblatt** vorliegen.
   b. Erste Zeile enthält **Variablennamen** (siehe Pkt. 3 oben!), ab der zweiten Zeile sind ausschließlich Daten enthalten.
   c. **Keine Farbcodierung** oder Kommentare, sondern ggf. zusätzliche Spalten einfügen, die die betreffende Information numerisch codiert enthalten.
   d. **Numerische Spalten** dürfen ausschließlich einzelne Zahlen enthalten (keine physikalischen Einheiten, keine „?“, keine Wertebereiche etc.). Es ist darauf zu achten, dass diese Spalten durchgehend als Zahlen formatiert sind.
   e. Keine Verwendung von Tausender-Trennzeichen.
   f. **Kategorielle Variablen** sind idealerweise numerisch kodiert. Im Falle von Texteingaben (z.B. „m“ bzw. „w“ beim Geschlecht) ist auf einheitliche Schreibweise zu achten.
   g. **Datumsvariablen** müssen durchgehend im selben Datumsformat eingegeben werden.
   h. Verwendung des durchgehend richtigen **Kommazeichens** beachten (Beistrich in deutscher, Punkt in englischer Version).
8. **Data Dictionary** (in MS Word oder MS Excel) als Ergänzung zum Datensatz
   a. kurze inhaltliche **Erklärung** der Spalten
   b. **Gliederung** in Hauptzielgröße(n), Nebenzielgrößen, Faktoren und Kovariablen, Adjustierungs- oder Stratifikationsvariablen
   c. Kennzeichnung der für die Studie **relevanten Variablen** (die restlichen Variablen müssen nicht gelöscht werden)
   d. Falls in SPSS keine Wertelabels verwendet wurden: Bedeutung der numerischen Codes für kategorielle Variablen

Sonstiges

9. **Archivierung der Daten**

   Für eine gesicherte Archivierung der Daten ist der/die DatenerstellerIn selbst verantwortlich.
10. **Aktualisierung von Daten**
    a. Nach Erhalt der Daten werden deskriptive Statistiken zur Plausibilitätsüberprüfung erstellt, um ev. Eingabefehler aufzudecken (**Data-Cleaning**).
    b. **Änderungen der Daten** nach Abschluss des Data-Cleanings sind zu vermeiden und führen ggf. zu Verzögerungen in der Auswertung. Bereits relativ einfache Analysen sind oft keinesfalls automatisch durchführbar, sondern bedürfen im Falle von Datenänderungen einer erneuten Schritt-für-Schritt-Vorgehensweise.
    c. In jedem Fall ist jede aktualisierte Version der Daten in derselben **Struktur** (Anordnung und Benennung der Spalten, Codierungen etc.) zu verfassen und mit Datum im Dateinamen zu versehen.
    d. Neu hinzu kommende Variablen können in einem eigenen Datensatz geliefert werden, der nur die Pat.identifikation und die neuen Variablen enthält.