Kleine Einführung in das
Data Management

© Dieter Augustin
Institut für Biometrie und Klinische Epidemiologie

Charité - Universitätsmedizin Berlin
CharitéCentrum 4 für Therapieforschung

Hinweis Dieses Skript setzt Grundkenntnisse in medizinischer Biometrie (Statistik) und Studienplanung voraus.
Im Skript werden gelegentlich geschützte Namen von Firmen und Produkten erwähnt, ohne ausdrücklich darauf hinzuweisen.

I. Data Management

1. Vorwort

Mit Themen wie Studienplanung und Data Management lassen sich ganze Bücher füllen. Dieses Skript soll einen kleinen Einblick in die Problematik und Tipps zur Lösung geben und Ihnen dabei helfen, Ihre Daten so zu erfassen, dass sie mit einem Statistikpaket wie *SPSS* ausgewertet werden können. Es wurde ursprünglich für unsere Einführungskurse in *SPSS* geschrieben, weshalb Beispiele sich meistens auf dieses Programm beziehen.

Aspekte wie die Besonderheiten multizentrischer und multinationaler Studien, industrielle Standardisierung, Kostenanalysen, Aufgaben eines Supervisors, Organisation von Fehlerreports, Schulung der Studienteilnehmer oder juristische Aspekte werden allenfalls am Rande erwähnt. Bei Bedarf lesen Sie bitte die Originalartikel bzw. weiterführende Literatur. Stichworte: ***SOPs*** = ***Standard Operating Procedures*** oder auf deutsch *Standardarbeitsanweisungen*, oft sehr umfangreiche Kataloge, die Schritt für Schritt das Vorgehen bei den diversen Arbeitsschritten wie Erstellung und Ausfüllen der Fragebögen, Datenerfassung oder Abschlussbericht festlegen. ***GCP*** = ***Good Clinical Practice*** bezeichnet das nicht immer offiziell festgelegte, aber allgemein übliche und etablierte Vorgehen bei klinischen Studien.

2. Einleitung

Eine medizinische Studie kann in die Abschnitte Planung, Datenerhebung, Auswertung und Abschlussbericht eingeteilt werden. In der Praxis lassen sich diese 4 Phasen nie strikt voneinander trennen. Oft sind Zwischenauswertungen sinnvoll oder sogar gefordert, während noch Daten erhoben und eingegeben werden. Manchmal erfordern die Ergebnisse eine nachträgliche Korrektur des Versuchsplans, was aber immer einen erheblichen Aufwand bedeutet, etwa wenn Daten nacherhoben werden müssen, um offen gebliebene Fragen zu klären. Außerdem müssen nachträgliche Änderungen (Amendments) einer angemeldeten Studie natürlich genehmigt werden.
Heute werden nicht nur die Daten für eine statistische Auswertung überwiegend mit Hilfe von Computern erfasst und ausgewertet. Auch für die Planung, das Schreiben von Berichten und die Erstellung von Grafiken wird EDV eingesetzt. Das erleichtert einerseits die Arbeit erheblich, erfordert aber andererseits, sich auf die Arbeitsweise von Computern einzustellen und die Bedienung der verschiedenen Programme zu erlernen. Murphys Gesetz **„If something can go wrong, it will“** (wenn etwas schief gehen kann, wird es das auch tun - und zwar im unpassendsten Moment) ist gewissermaßen der Leitsatz für den Umgang mit Rechnern. Ein Computer ist eine Maschine, die alle Anweisungen stur befolgt. Außerdem gehören Schleifen zu jedem Programm. Bei größeren Datenmengen oder rekursiven Formeln kommen schnell ein paar hundert oder tausend Durchläufe zusammen. Da wächst auch die Wahrscheinlichkeit, dass ein seltener Fehler irgendwann auftritt. Ein Beispiel, das Schlagzeilen machte, waren die ersten Pentiums mit einem Fehler in der FPU. Obwohl die Fehlerwahrscheinlichkeit laut Hersteller winzig war, kam es weltweit immer wieder zu falschen Ergebnissen und *INTEL* musste die fehlerhaften Prozessoren schließlich austauschen.

Die Anforderungen an das Data Management sind inzwischen sehr hoch. Vor allem bei Studien, in denen es um viel Geld geht, etwa die Zulassung eines neuen Medikaments, ist die Versuchung groß, das Ergebnis mit kleinen „Datenkorrekturen“ zu „verbessern“. Solche Studien werden strenger überwacht als die Bilanzen einer Firma, und jede nicht exakt nachgewiesene und begründete Änderung an den Daten kann zur Folge haben, dass die ganze Studie abgelehnt wird. Bei solchen Projekten sollten aber unbedingt Spezialisten einbezogen werden. Dieses Skript soll Ihnen helfen, Ihre Daten computergerecht zu erheben und zu erfassen, damit eine Auswertung mit gängigen Statistikpaketen ohne große Probleme möglich ist. Grundlage sind zwei Artikel aus einem Sonderheft der Zeitschrift *„Controlled Clinical Trials“* ♦.

♦ **Controlled Clinical Trials**,
Volume 16, Number 2, 1995
Supplement „Data Management for Multicenter Studies: Methods and Guidlines“.
McFadden, Eleanor T., LoPresti, Francet, Bailey, Lance R. et al: Approaches to Data Management
James D. Hosking, M. Marvin Newhouse, Anna Bagniewska et al: Data Collection and Transcription.

3. Versuchsplanung

Eine gute prospektive Planung ist das A und O und kann viel Zeit, Kosten und Ärger ersparen. Je besser die Planung vor Studienbeginn ist, um so besser wird die Studie sein. Dies wird umso wichtiger, je größer das Projekt ist (große Datenmengen, langer Zeitraum, multizentrisch), aber auch bei kleineren Arbeiten wie Dissertationen kann eine unzureichende und schlechte Planung katastrophale Folgen haben und die ganze Arbeit in Frage stellen.
Als erstes muss das wissenschaftliche Konzept entwickelt werden. In einem Protokoll sind das Ziel und die Fragestellung der Studie sowie der zu begehende Weg festzulegen, wobei bereits die Aspekte des Data Managements zu berücksichtigen sind. Ein vollständiger, konsistenter Plan ist entscheidend für das Data Management System und letztendlich für den Erfolg der Studie insgesamt. Meistens muss ein konkreter und korrekter Prüfplan erstellt werden, bevor mit der Studie begonnen werden kann, sei es zur Genehmigung (Ethik-Kommission) oder zur Vorlage beim Sponsor. Das einmal beschlossene und genehmigte Protokoll ist verbindlich und darf eigenmächtig nicht mehr geändert werden.

Folgende Punkte müssen vor Studienbeginn bedacht werden:

- **Das Ziel der Studie (Fragestellung)**
- **Das Studienpersonal**
  Werden Fachleute gebraucht (Vollzeit oder als Berater)? Können Hilfskräfte für Routinearbeiten (z.B. Datenerfassung) eingesetzt werden?
- **Die Fallzahl**
  Wie viele Fälle Sie für eine Studie einplanen müssen, lässt sich nie pauschal beantworten. Für ein immer hochwirksames Mittel benötigen Sie nur wenige Testpersonen, ein unwirksames Medikament wird auch bei zahlreichen Probanden allenfalls ein zufälliges Ergebnis zeigen. Für eine sinnvolle Fallzahlschätzung müssen Sie nicht nur die Irrtumswahrscheinlichkeit α festlegen, sondern auch den Fehler 2. Art bzw. die Power (1-ß) sowie den Unterschied zwischen den Gruppen und die Streuung abschätzen können.
  Im Prinzip müssen Sie in den statistischen Test, den Sie am Ende rechnen wollen, die erwarteten Ergebnisse einsetzen und die Formel nach n auflösen. Das ist in der Praxis meistens nicht möglich, so dass Näherungsverfahren verwendet werden müssen. Es ist nicht zu empfehlen, die Fallzahlberechnung auf eigene Faust zu versuchen. Es gibt wenige (teure) Programme, die allgemein anerkannt sind, etwa ***nQuery*** oder das Statistikprogramm ***SAS***. Lassen Sie sich beraten.
- **Die Probanden**
  Wie können Probanden für die Studie rekrutiert werden? Handelt es sich um „normale“ Patienten, die vielleicht etwas intensiver beobachtet werden, oder müssen die Probanden eingeworben werden? Wie aufwendig oder unangenehm ist die Studienteilnahme für sie? Muss eine Aufwandsentschädigung gezahlt werden? Mit wie vielen Fällen kann gerechnet werden, wie lange dauert es, die errechnete Fallzahl zu erreichen? Juristische Aspekte (Aufklärung, Einwilligung, Ethik-Kommission, Versicherungen)?
- **Der zeitliche Ablauf und der Datenfluss**
  Wie lange dauert die Erhebung für einen Patienten? In welchem Zeitabstand sind Nachuntersuchungen nötig? Steht die benötigte Probandenzahl bei Studienbeginn zur Verfügung, oder werden sie nach und nach aufgenommen? Wie kommen Sie zu den Daten (Fragebögen, Telefonumfrage, Ärzte, Akten)? Wie lange werden neue Probanden aufgenommen? Wann ist die Studie zu Ende?
- **Die Struktur der Daten**
  Welche Größen erheben? Wie viele Daten pro Proband? Wie groß wird die gesamte Datenmenge? Wie codieren? Wie sieht der Fragebogen aus?
- **Die Datenerfassung**
  Wer gibt die Daten in den Computer ein? Mit welchen Programmen? Ist eine Datenbank erforderlich oder genügt eine einfache Datenmatrix? Welche Hard- und Software wird benötigt?
- **Die Auswertung**
  Welche statistischen Verfahren? Reicht ein übliches Statistikpaket oder brauche ich spezielle Software? Müssen neue Programme entwickelt werden?

3.1. Das Studienpersonal

Inhaltliche Fragen kann oft nur ein Mediziner klären, meistens Spezialist auf dem entsprechenden Fachgebiet und in der Regel Initiator der Studie. Er muss u.a. bestimmen, welche Merkmale für die spezifische Fragestellung wichtig sind und mit welchen Störgrößen und Nebenwirkungen zu rechnen ist. Außerdem sollte er die Ergebnisse kritisch betrachten: sind sie medizinisch relevant?
Der Datenmanager muss dafür sorgen, dass die Daten vollständig und in computergerechter Form erhoben und so erfasst werden, dass sie für die Auswertung brauchbar sind.
Der Biometriker hat zu überwachen, dass Verfahren angewendet werden, die formal korrekt und zur Beantwortung der Fragestellung geeignet sind.

Dazu können noch Laboratorien (Untersuchungen), die Klinikumsapotheke (wichtiger Partner bei doppelt blinden Studien) und andere Einrichtungen kommen.
In der Praxis werden oft mehrere oder auch alle diese Funktionen von einer Person ausgeführt. So gehört der Umgang mit Computern zum Alltag der Biometriker, die also oft auch die Rolle des EDV-Spezialisten übernehmen können. Bei kleineren Arbeiten wie Dissertationen kann und muss der Doktorand alles alleine machen, abgesehen vielleicht von einer statistischen Beratung. Bei komplexen Problemen, die viel Hintergrundwissen erfordern, sind Spezialisten dagegen unabdingbar. So gab es z.B. in den 80er Jahren ein Projekt zur numerischen Auswertung von kranialen Computertomographiebildern, die ohne einen Programmierer (decodieren der Originalbänder, Umsetzen auf das Format des FU-Rechners, diverse Auswertprogramme), einen Biometriker und Mathematiker (Verfahren, um die gewaltige Datenflut sinnvoll zusammenzufassen und verschiedene Bilder zu vergleichen) und natürlich einen Gehirnspezialisten (Initiator der Studie, verantwortlich für alle medizinischen Fragen) nicht möglich gewesen wäre.
Oft müssen also Personen aus verschiedenen Fachrichtungen zusammenarbeiten, um eine Studie durchzuführen. Wichtig ist in diesen Fällen, dass sich die jeweils Verantwortlichen schon in der Planungsphase zusammensetzen. Andernfalls passiert es immer wieder, dass sie in der vorhandenen Form nicht durchführbar ist oder die bereits erhobenen Daten nicht zu gebrauchen sind. Die Folge ist im günstigsten Fall ein großer zusätzlicher Aufwand, um die gemachten Fehler zu korrigieren, im ungünstigsten Fall muss die Studie verworfen oder ganz neu gestartet werden.

3.2. Verblindung und Randomisierung

3.2.1. Verblindung

Häufig werden Studien „blind“ oder „doppelt blind“ durchgeführt. Geben Sie z.B. Patienten zur Beruhigung Dragees mit Johanniskraut und befragen sie anschließend nach der Wirkung, so könnte ein positives Urteil durchaus als Placebo-Effekt angesehen werden. Um das auszuschließen, bekommt die Hälfte der Patienten wirklich Johanniskraut und die andere Hälfte Placebo (Dragees ohne jeden Wirkstoff). Dann müsste sich bei Wirksamkeit ein Unterschied zwischen der Placebo- und der Verumgruppe feststellen lassen. Es macht aber keinen Sinn, wenn die Probanden wissen, dass sie ein wirkstofffreies Medikament schlucken, d.h. sie erfahren nicht, zu welcher Gruppe sie gehören („blind“). Dieses Prinzip lässt sich steigern, indem auch der behandelnde Arzt nicht weiß, was er dem Probanden gibt („doppelt blind“). Das schützt ihn vor ungewollter Subjektivität oder auch nur davor, dass ihm Subjektivität vorgeworfen wird. Auch die Auswertung kann so durchgeführt werden, dass der Biometriker zwar weiß, welcher Proband in welche Gruppe fällt, aber nicht was die Gruppen bedeuten (1=Placebo und 2=Verum oder ist es umgekehrt?)
In der Regel versucht man, Studien immer doppelt blind durchzuführen. Manchmal ist das aber nicht praktikabel, etwa bei einer Akupunkturstudie.

3.2.2. Randomisierung

Am sinnvollsten ist es, die Patienten zufällig auf die Gruppen zu verteilen, um jeder Systematik vorzubeugen. Zu viel Zufall kann aber schädlich sein und die ganze Studie unbrauchbar machen. Im Prinzip könnte man die Zuordnung mit einer Münze durchführen (Kopf = Placebo, Zahl = Verum). Nehmen wir an, in jeder Gruppe sollen 10 Probanden sein, insgesamt also 20. Mit einer Münze kann es leicht passieren, das z.B. 14 Probanden in die Gruppe 1 kommen und nur 6 in Gruppe 2. Am Ende bekommt man trotz des vorhandenen Effektes kein signifikantes Ergebnis, weil die Fallzahl in der einen Gruppe zu klein ist. Wenn man sich schon bei der Planung die Mühe einer Fallzahlschätzung gemacht hat, soll die Randomisierung sich auch an diese Planung halten! Hier würde man lieber 20 Lose herstellen und den Probanden zuordnen.
In der Regel verwendet man heute Programme zur Randomisierung. Das ist eigentlich widersprüchlich, weil es bei Computern keinen Zufall gibt. Deshalb spricht man auch von Pseudo-Zufallszahlenprogrammen. Diese erzeugen meistens nach einem bestimmten Algorithmus gleichverteilte Zahlen zwischen null und eins, wobei die zuletzt errechnete Zahl als Basis für die nächste dient. Erzeugt man mehrmals 100 „Zufallszahlen“ mit dem gleichen Startwert (auch *Seed* genannt), so bekommt man immer das gleiche Ergebnis. Bevor man Zufallszahlen erzeugt, gibt man also einen willkürlichen Startwert ein; viele Programme benutzen dafür die Sekundenbruchteile aus der Systemuhr. Manche Studien verlangen, dass auch die Randomisierung nachvollziehbar ist. Dann muss man den benutzten Startwert unbedingt speichern.
In unserem Beispiel würde man einen Startwert setzen und dann 20 Zufallszahlen erzeugen. Die Patienten mit den 10 kleinsten Zahlen kommen in Gruppe 1, der Rest in Gruppe 2.

3.2.2.1. Blockung

Bei einer großen Fallzahl könnte es passieren, dass zu Beginn der Studie eine Gruppe gehäuft auftritt. Auch das ist problematisch. Man könnte z.B. vermuten, dass die Ärzte im Laufe der Studie routinierter werden und sich dadurch der Erfolg steigert. Ein signifikanter Unterschied zwischen den Studiengruppen könnte so zustande kommen und nicht, weil eine Methode besser ist als die andere. Und falls eine Studie vorzeitig abgebrochen wird, ergibt sich wieder eine Schieflage bezüglich der Fallzahl.

Um diesem Problem zu entgehen, randomisiert man größere Fallzahlen nicht auf einmal, sondern in kleinen „Portionen“. Bei 120 Probanden könnte man z.B. 15 Blöcke zu je 8 Patienten randomisieren. Dann geht nach jeweils 8 Probanden die Gruppenaufteilung auf und eine extrem schiefe Verteilung ist nicht zu befürchten.

3.2.2.2. Schichtung

Manchmal sind an einer Studie mehrere Zentren oder Ärzte beteiligt. Auch hier möchte man vermeiden, dass die Probanden ungleichmäßig verteilt werden und signifikante Unterschiede dem erfahreneren Arzt oder dem besser ausgestatteten Klinikum zugeschrieben werden könnten. Es kann auch andere Gründe für eine Schichtung geben. Es ist z.B. bekannt, dass der „Altersdiabetes“ immer häufiger auch bei jungen Menschen diagnostiziert wird. Da die Möglichkeit besteht, dass die jüngeren Patienten anders auf eine Therapie reagieren, kann man auch hier zwei Schichten anlegen. In der Praxis erzeugt man einfach für jede Schicht eine separate Randomliste. Dann kann man bei genügender Fallzahl auch die Schichten einzeln auswerten (z.B. nur die älteren Diabetiker) und die Randomisierung ist balanciert.

3.2.3. Durchführung einer verblindeten randomisierten Studie.

Die Randomlisten werden von einer Person erstellt, die nicht an der praktischen Durchführung der Studie beteiligt ist und sie natürlich unter Verschluss hält. Nur die Apotheke bekommt die Liste ausgehändigt und erstellt die entsprechenden Präparate, die für beide Gruppen nicht zu unterscheiden sind. Die Gebinde werden nur mit der Nummer des Probanden beschriftet. Der erste Patient, der im Zentrum 3 aufgenommen wird, bekommt z.B. die Nummer 3001 zugewiesen und entsprechend die Präparate mit dieser Beschriftung.

3.2.3.1. Notfallumschläge

Der Arzt soll nicht wissen, was er dem Patienten verabreicht. Es kann aber zu Notfällen kommen, in denen er aus medizinischen Gründen den „Code brechen“ muss. Auf keinen Fall darf der Arzt dann die Randomliste einsehen, da sonst die Studie beendet wäre. Deshalb hinterlegt man für jeden Probanden einen verschlossenen Umschlag, der im Notfall geöffnet werden darf. Diese Umschläge dürfen auf keinen Fall weggeworfen werden, sondern sie sind am Ende der Studie zurückzugeben als Beweis, dass der Umschlag (und damit das Studienprotokoll) nicht verletzt wurde. Andernfalls muss in der Regel ein spezieller CRF ausgefüllt werden, um die Öffnung des Umschlags zu begründen (z.B. schwerwiegende Nebenwirkungen).

4. Grundlagen des Data Managements

Wenn einem Arzt irgendwelche Besonderheiten auffallen, die er gerne mit Hilfe einer klinischen Studie untersuchen möchte, interessieren ihn eigentlich nur die Fragestellung und die Ergebnisse. Damit aber brauchbare Ergebnisse herauskommen, müssen Daten in einer geeigneten Form gewonnen und archiviert werden. Alles, was mit der Verwaltung der Daten zu tun hat, fasst man unter dem Sammelbegriff ***Data Management*** zusammen. Dazu gehört nicht nur die computergerechte Codierung und Erfassung der Daten, sondern auch z.B. die Ablage der Originalbelege. Wer schon einmal einen bestimmten Fragebogen gesucht hat (z.B. weil die Daten auf dem PC offensichtlich fehlerhaft waren) und dafür 20 volle Ordner durchblättern musste, weiß, dass auch die Fragebögen möglichst sortiert abzuheften sind.
Das Ziel des Data Managements ist die Vorbereitung der Zwischen- und Endauswertung. Die Veröffentlichung (als Artikel, Vortrag, Dissertation oder Habilitation) stellt in der Wissenschaft das „Endprodukt" dar. Studienplanung und -koordination, Datenerhebung und -auswertung, Qualitätskontrolle, Zwischenberichte und Analysen dienen dazu, dass die veröffentlichten Ergebnisse akkurat und valide sind.

In den meisten Projekten wird mehr Zeit und Personal für die Datengewinnung und -erfassung benötigt als später für die eigentliche Auswertung. Die Schritte sind:

- Daten erheben und festhalten; meistens auf Papier (Fragebogen), manchmal auch direkt auf einem elektronischen Medium (Notebook, USB-Speicher).
- Codierung oder Klassifizierung von Informationen, die in der vorliegenden Form nicht sinnvoll von einem Computer verarbeitet werden können, z.B. freier Text, Röntgenbilder, EKGs oder Tonaufnahmen.
- Umwandeln der Information auf Papier in elektronisch lesbare Form, also Eingabe in den Computer; meistens über die Tastatur.
- Überprüfen der Daten auf Unstimmigkeiten und offensichtliche Fehler, ggf. Korrektur.

5. Datenerhebung

5.1. Allgemeines

Die Datenerhebung ist eine wichtige und kritische Phase. Falsch gemessene Werte lassen sich später nicht mehr korrigieren und sind häufig auch nicht zu erkennen. Auch das beste Datenbanksystem und der schlaueste Mathematiker können falsch erhobene Größen nicht mehr „richtig rechnen".
Die erste Frage ist, welche Daten zu welchen Zeitpunkten bzw. in welchen Intervallen benötigt werden. Die meisten Items werden durch die Fragestellung von dem verantwortlichen Mediziner festgelegt, einige zusätzliche Variablen sind zur Kontrolle des Datenflusses unbedingt notwendig.
Es sollte gut überlegt werden, was tatsächlich nötig ist. Welche Daten gesammelt werden, richtet sich nach der Fragestellung der Studie und nicht nach den klinischen Notwendigkeiten. Es kann durchaus sein, dass aus klinischer Sicht große Datenmengen gesammelt werden, damit der Arzt auf Abweichungen sofort reagieren kann (z.B. Patienten auf der Intensivstation, deren EKG ständig aufgezeichnet wird). Es ist oft unnötig, jeden Arztbesuch und jede Untersuchung in eine Studie aufzunehmen. Ein übermäßiges Anwachsen des Datensatzes (gemeint ist vor allem die Zahl der Merkmale, nicht die Fallzahl) bedeutet nicht nur einen größeren Aufwand bei der Datenerhebung und -eingabe. Meistens geht auch die Übersicht über die Daten verloren, was zu Fehlern in der Erhebungsphase und oft auch zu konfusen Auswertungen führt. Andererseits können vergessene Werte die Beantwortung der Fragestellung unmöglich machen. Außerdem dürfen Sie nicht nur die Merkmale erfassen, die Sie unmittelbar betrachten wollen, sondern müssen auch überlegen, welche *Confounder* (Störgrößen) Ihre Ergebnisse beeinflussen können (in der Medizin sind das u.a. fast immer Alter und Geschlecht).
Bedenken Sie, dass Sie nur vergleichbare Daten statistisch auswerten können. Das bedeutet nicht nur, dass Sie für alle Messungen eines Parameters die gleichen Einheiten verwenden und gleiche Fragen immer auf die gleiche Art codieren. Genauso wichtig ist die inhaltliche (logische) Vergleichbarkeit der Werte.

5.2. Die Datenmatrix

Damit die Auswertung der Daten durch ein Statistikpaket wie SPSS möglich wird, müssen die Daten immer als rechteckige Tabelle (Matrix) vorliegen. Dabei bilden die Fälle (Probanden, Merkmalsträger) die Zeilen und die beobachteten Merkmale die Spalten. Beispiel:

	LfdNr	Alter (Jahre)	Geschlecht (1=männlich, 2=weiblich)	Größe (Meter)	Gewicht (kg)	...
Fall 1	1	36	2	1,71	69	...
Fall 2	2	23	1	1,85	91	...
Fall 3	3	50	1	1,69	70	...
Fall ...	...	...	...	...	...	...

- Zahl und Bedeutung der erfassten Merkmale (Variablen) sind für jeden Fall gleich, nur die Merkmalsausprägungen unterscheiden sich.
- In jede Zelle (= Schnittfeld zwischen Fall und Merkmal) passt ***exakt ein Wert*** (eine Aussage). Die möglichen Ausprägungen für diese Zelle müssen ***disjunkt*** (eindeutig, ohne Überschneidungen) und ***vollständig*** sein. Mehrfachantworten oder Aufzählungen sind nicht möglich!
- Die Datenmatrix muss immer vollständig ausgefüllt sein, d.h. in jeder Zelle muss ein Wert stehen. Ggf. sind Antwortmöglichkeiten wie „*sonstige*", „*weiß nicht*", „*keine Angabe*" oder „*not done*" vorzusehen. Besonders bei Daten aus Medizin und Biologie ist immer damit zu rechnen, dass manchmal keine Antworten existieren (eine Untersuchung wurde nicht durchgeführt, der Patient konnte oder wollte nicht befragt werden, ...). Um solche *fehlenden Angaben* einzugeben, werden entsprechende Codes eingetragen (s. Abschnitt *Fehlende Werte*).

Diese rechteckige Datenstruktur sollte von Anfang an berücksichtigt werden, sonst wird die statistische Auswertung erheblich erschwert oder gar unmöglich. Es ist nötig, gewisse Formalismen strikt einzuhalten, damit eine maschinelle Auswertung per Computer erfolgen kann. Das zugrunde liegende Prinzip ergibt sich aber aus den Anforderungen für eine sinnvolle statistische Analyse. Je besser eine Studie geplant und je konkreter die Fragestellung formuliert ist, umso leichter wird die Erstellung einer korrekten Datenmatrix.

5.2.1. Eine Datenspalte

Alle Werte in einer Spalte sind verschiedene Ausprägungen des gleichen Merkmals bei unterschiedlichen Personen. Damit sie vergleichbar sind, müssen sie unter möglichst gleichen Bedingungen erhoben werden, d.h. alle Faktoren, die die Messung bzw. Beobachtung beeinflussen können, sollten möglichst gleich sein (mit Ausnahme der zu untersuchenden Einflussgrößen). Dazu gehört oft auch der Zeitpunkt der Beobachtung. Wenn Sie in eine Spalte z.B. die mit Ultraschall ermittelte Größe eines Feten eingeben, sollten die

Daten alle aus der gleichen Schwangerschaftswoche stammen. Es dürfte wenig sinnvoll sein, Messungen aus der zehnten und aus der zwanzigsten SSW zusammenzuwerfen.
Im Idealfall erheben Sie Ihre Daten nach einem vorgegebenen Plan. Dann ist es kein Problem, die Daten als Matrix darzustellen. Häufig hat man aber mehr Daten, als wirklich benötigt werden. Dann muss versucht werden, diese zu verdichten. Auf welche Weise das geschehen kann, bestimmt in erster Linie die Fragestellung der Studie, also der verantwortliche Arzt. Hier ein paar typische Beispiele:

- Wenn **gleichwertige Messungen** vorliegen, können Sie diese durch geeignete Maßzahlen (Median, Mittelwert, Minimum, Maximum etc.) ersetzen. Beispiel: Von bestimmten Personen wurde im Laufe der Zeit mehrfach der Cholesterinwert bestimmt. Die Zeitpunkte sind eher zufällig und an kein bestimmtes Ereignis gebunden. Durch die Bildung eines Mittelwertes werden zufällige Extremwerte geglättet.
- Bei **Zeitreihen** legen Sie die Untersuchungstermine im Voraus fest. Der Zeitpunkt Null ist immer ein konkretes Ereignis, z.B. eine Operation oder der akute Ausbruch einer Erkrankung. Die Anzahl und die Abstände der Untersuchungen richten sich nach den medizinischen Gegebenheiten, ebenso wie die noch tolerierbaren zeitlichen Abweichungen (z.B. 1 Tag vor Operation, direkt nach der Op, nach 1 Woche, nach 2 Wochen, nach 4 Wochen, nach 2 Monaten und nach 6 Monaten). Basis sollte in jedem Fall die „Stunde Null“ sein und nicht der jeweils vorangehende Termin, um die Kumulation von Zeitverschiebungen zu vermeiden. Im Beispiel ist der 5. Stichtag also 4 Wochen nach der Operation und nicht etwa 2 Wochen nach der 4. Untersuchung.
  Falls es sich um eine retrospektive Studie handelt, die Daten also schon existieren, müssen die passenden Zeitpunkte herausgesucht und die entsprechenden Werte eingegeben werden. Je kleiner Sie die erlaubten Abweichungen vom Sollzeitpunkt ansetzen, umso mehr fehlende Werte werden Sie erhalten. Eine ungerechtfertigte Großzügigkeit kann zur Verfälschung der Daten führen. In jedem Fall ist es in der Praxis meistens einfacher, die Daten von Hand herauszusuchen, als erst ein Computerprogramm zu diesem Zweck zu entwickeln und auszutesten.
- Das oben Gesagte gilt im Prinzip auch dann, wenn nicht der zeitliche Verlauf, sondern der **Zeitpunkt** für die Studie wichtig ist (Beispiele: Werden bei Vollmond mehr Kinder geboren? Ist die Unfallhäufigkeit montags am größten? Sind Erkältungen im November oder im April häufiger?).
- Wenn es um das **Auftreten eines bestimmten Ereignisses** geht, genügt oft die Zeit, die zwischen der „Stunde Null“ und dem Ereignis verstrichen ist (z.B. das Auftreten einer bestimmten Komplikation nach einer Operation). Eine weitere Möglichkeit wäre festzustellen, wie oft das Ereignis in einem bestimmten Zeitraum auftritt (z.B. Zahl der grippalen Infekte in einem Winter, Extrasystolen im 24-h-EKG).

5.2.2. Verbundene Stichproben

Alle Werte innerhalb einer Datenzeile gehören zum gleichen Probanden. Bei einem verbundenen Test werden individuelle Änderungen eines Merkmals betrachtet (z.B. Körpergewicht vor und nach einer Diät). Folglich werden verbundene Tests immer zwischen zwei Spalten (Variablen) berechnet. Welcher Test sinnvoll auf welche Variablen anzuwenden ist, entscheidet ausschließlich der Anwender. In SPSS lautet der Befehl für den Wilcoxon-Test

LFN	GRUPPE	KGW0	KGW1
1	1	75	73
2	2	88	79
3	2	69	71
4	1	93	88
5	2	71	71

NPAR TESTS WILCOXON = KGW0 WITH KGW1.

5.2.3. Unverbundene Stichproben

LFN	GRUPPE	KGW0	KGW1
1	1	75	73
2	2	88	79
3	2	69	71
4	1	93	88
5	2	71	71

Bei unverbundenen Stichproben handelt es sich um verschiedene Individuen, die Sie anhand eines bestimmten Merkmals in Gruppen zusammenfassen können (Frauen / Männer; Diabetiker / Gesunde). Sie müssen also die Fälle (Zeilen) in der Datenmatrix in zwei oder mehr Gruppen aufteilen. Das ist kein Problem, wenn die benötigten Werte in der Datenmatrix vorhanden sind. *Beispiel:* Der Befehl

NPAR TESTS M-W = KGW1 BY GRUPPE (1,2)

teilt den Datensatz in die beiden Gruppen GRUPPE=1 und GRUPPE=2 und führt damit einen Mann-Whitney-Test über die Variable KGW1 durch.

5.3. Die Bezugsgröße

Wer bzw. was ist der Merkmalsträger, d.h. die Versuchseinheit? Normalerweise ist das immer das Individuum, also der Patient ober Proband. Ausnahmen von dieser Regel sollten sehr gut überlegt werden, da sie das Studienergebnis verfälschen können. So könnten Laboruntersuchungen - wenn die Einheit die Untersuchung (der Laborbogen) ist und nicht der Patient - zu schlechteren Ergebnissen führen, da z.B. das Blut von kritischen Patienten auf der Intensivstation erheblich häufiger untersucht wird als bei weniger ernst erkrankten Personen.

Es kann manchmal sinnvoll sein, als Einheit z.B. ein Auge, ein Knie oder ein Zahnimplantat zu nehmen. Aber beide Augen eines Patienten unterliegen dem gleichen Stoffwechsel und sind somit nicht wirklich unabhängig. Stößt das linke Auge eine künstliche Linse ab, dürfte am rechten das gleiche passieren. Steht dagegen die Technik oder das Geschick des Operateurs im Vordergrund, so kann man durchaus die Operation als Bezugsgröße nehmen, zumal in der Regel nie beide Augen gleichzeitig operiert werden. In einer soziologischen Untersuchung könnte die Beobachtungseinheit aber auch eine ganze Familie oder eine Firma sein.

6. Der Dokumentationsbogen (CRF)

Die Abkürzung ***CRF*** kann viele Bedeutungen haben. In diesem Zusammenhang steht sie für ***Case Report Form*** und wird oft als Synonym für Fragebogen bzw. Dokumentationsbogen verwendet.
Das Ziel beim Entwurf eines Dokumentationsbogens bzw. Fragebogens ist ein Instrument zur Datengewinnung und -erfassung mit den Eigenschaften

- Relevant in Bezug auf das Ziel der Studie
- Akkurat und vollständig
- Standardisiert für alle an der Studie beteiligten Zentren (innerhalb der Studie)
- Standardisiert in Bezug auf ähnliche Studien (zwischen Studien)
- Effizient für die Datenerhebung
- Effizient für die Datenerfassung
- Geeignet für die statistische Analyse.

In seltenen Fällen kann es zwischen den oben aufgeführten Kriterien zu Konflikten kommen, meistens werden sie sich aber gegenseitig ergänzen. Ein Fragebogen, bei dem die gegebenen Antworten schnell auszulesen sind (z.B. Kästchen am rechten Rand) erleichtert nicht nur die Datenerfassung auf einem Computer, er ist übersichtlicher und einfacher auszufüllen und Lücken fallen auf.
Die für die Eingabe in den Computer wichtigen Daten müssen leicht zu erkennen und richtig sein. Das „Mitdenken“ bei der Eingabe kostet Zeit und führt leicht zu Fehlern. Ausnahmen wie „wenn in Frage 18 eine ‚4‘ steht, muss statt dessen eine ‚7‘ eingetippt werden“ werden in der Routine schnell vergessen. Bedenken Sie auch, dass zur Erfassung oft Hilfskräfte eingesetzt werden, denen die medizinische Terminologie nicht geläufig ist und die mit schlecht lesbarer Schrift große Probleme haben!
Es kann durchaus sinnvoll sein, Fragebögen aus ähnlichen Studien als Vorlagen zu benutzen. Das erspart nicht nur Arbeit beim Entwurf, sondern erleichtert den Vergleich der eigenen Ergebnisse mit denen der anderen Studie. Allerdings darf auch hier die kritische Überprüfung des Bogens nicht entfallen; es hat wenig Sinn, eventuell die gleichen Fehler wie seine Vorgänger zu machen.
Meistens läuft die Entwicklung eines Fragebogens so ab, dass ein Mitarbeiter der Studie (z.B. mit einem Textverarbeitungsprogramm) die Formulare entwirft und ausdruckt. Diese werden dann besprochen, korrigiert, neu erstellt und abermals besprochen, bis ein Konsens aller Beteiligten erreicht ist. Zuerst sollten allerdings die benötigten Items in einem Fragenkatalog zusammengestellt werden, das Layout des Bogens ist erst der letzte Schritt. Sonst ist die Gefahr groß, dass bei jeder Sitzung die Fragebögen völlig umgestaltet werden und am Ende aus Zeitmangel ein schlechter Fragebogen verwendet wird.
In vielen Studien dient der Fragebogen zur Erfassung der Originaldaten, etwa wenn bei einem Interview die Antworten eingetragen werden. Es versteht sich von selbst, dass hier eine ganz besondere Sorgfalt gefordert ist, da eine nachträgliche Korrektur praktisch unmöglich ist. Die ausgefüllten Bögen sind wichtige Dokumente und entsprechend zu behandeln und aufzubewahren. Sammeln Sie alle Unterlagen von einem Probanden (Einwilligung, Fragebögen, Laborbögen, Tagebücher etc.) in einer Mappe und diese Mappen wiederum nach Patientencode sortiert in Ordnern, Schubern oder einer Registratur. Beschaffen Sie vor Beginn der Studie den benötigten Platz und die Materialien, damit jeder Beleg an seinem Platz aufbewahrt und im Zweifelsfall schnell wiedergefunden werden kann.

6.1. Identifikatoren

Auch wenn die Zeit der Lochkarten vorbei ist: Jedes Formular (selbst jedes einzelne Blatt darin), jeder Fall und jeder Datensatz im Computer müssen ohne großen Aufwand eindeutig zu identifizieren sein. Ein kurzer Blick auf das Deckblatt sollte genügen, um z.B. festzustellen, dass es sich um den Laborbogen der zweiten Untersuchung für Patient Nummer 3251 für die Studie xyz handelt. Alle Folgeblätter sollten diese Informationen ebenfalls beinhalten, um ggf. lose Blätter wieder korrekt einzuordnen. Wenn Sie die Bögen mit einem Textprogramm wie Word erstellen, können Sie Studie, Fragebogen, Patientennummer und eine Seitennummerierung in den Kopf- oder Fußzeilen unterbringen. Dann müssen Sie allerdings alle Bögen über den Drucker ausgeben und können keinen Kopierer benutzen. Die Summe aller für die Identifizierung benötigten Felder bezeichnet man als Schlüssel oder Index. Jedes in einer Studie benutze Formular muss einen solchen Schlüssel haben und die Schlüsselfelder sollten für alle Formulare in Art und Anordnung gleich sein.
Man unterscheidet typischerweise 4 verschiedene Identifikatoren: Für die Studie, die benutzten Formulare, die Teilnehmer (Probanden) und für die erfassten Merkmale (Items).

6.1.1. Identifikatoren für die Studie

Auf jedem Fragebogen sollte vermerkt sein, zu welcher Studie (welchem Projekt) er gehört. Der Name sollte eindeutig sein, also z.B. nicht „Kariesstudie“, sondern „Kariesprophylaxe 2006 Zahnklinik Charité Berlin“. Wenn es eine offizielle Studiennummer gibt, sollte diese immer vermerkt werden. Im Computer muss der Name nicht in jeden Datensatz eingegeben werden, wenn alle Daten, die zu der Studie gehören, physikalisch zusammengefasst werden (eigene Platte bzw. bei kleineren Projekten eigenes Verzeichnis). Dies gilt nicht,

wenn die Datensätze dieser und einer anderen Studie zusammengemischt werden sollen, dann ist ein Identifyer-Feld für jeden Datensatz nötig (das möglichst automatisch eingefügt wird)!

6.1.2. Identifikatoren für die Teilnehmer

Jeder Teilnehmer muss eindeutig identifizierbar sein. Das heißt, dass nie zwei Probanden den gleichen Code haben dürfen, dass aber andererseits auch für eine Person nie mehrere Codes existieren dürfen. Dieser Identifikator gilt für alle in der Studie verwendeten Fragebögen und ist - besonders beim EDV-Einsatz - der Schlüssel, um Daten für eine Person aus mehreren Fragebögen zusammenzufügen.
Zu unterscheiden sind „natürliche“ und „zugewiesene“ Identifikatoren. Zu den natürlichen zählen Angaben wie Name, Alter, Geschlecht usw. In Krankenhäusern werden oft mehrere dieser Angaben zusammengesetzt, um daraus eine I-Zahl herzustellen. Dieses Verfahren ist aber nicht absolut eindeutig, es kann immer wieder zufällig gleiche I-Zahlen bei völlig verschiedenen Patienten geben. Was für die Ablage von Krankenakten durchaus sinnvoll ist (die I-Zahl lässt sich einfach rekonstruieren, ohne dass der Patient eine Nummer lernen muss; die Krankenakte lässt sich bei Wiederaufnahme leicht finden), ist für Studien nicht praktikabel.
Bei Studien ist es üblich, einfach jedem Probanden eine laufende Nummer zuzuweisen, bei randomisierten Blindstudien geht es ohnehin nicht anders. Manchmal ist es sinnvoll, zusätzliche Information in dieser Nummer zu „verstecken“, etwa die Zugehörigkeit zu einer bestimmten Schicht. Dann muss diese aber immer in den höchsten Stellen stehen und nicht in den Einerstellen, damit ein numerisches Sortieren möglich wird. In einer multizentrischen Studie bekommt z.B. jedes Zentrum einen Code, der in der Patientennummer als Tausenderstelle erscheint. Dann kann jedes Zentrum seine Fälle mit einer lückenlosen laufenden Nummer versehen (wenn es mehr als 9 Zentren oder mehr als 999 Fälle in einem Zentrum gibt, müssen die Stellen entsprechend verschoben werden). Dann ist sofort erkennbar, dass Patient 3045 der 45. Fall aus dem Zentrum mit der Nummer 3 ist. Natürliche Identifikatoren können zusätzlich aufgenommen werden, um Fehler bei der Zuordnung zu erkennen, z.B. ein 5stelliger Textcode aus den beiden ersten Buchstaben des Vor- und den drei ersten des Nachnamens. Dieses Verfahren ist dringend empfohlen, wenn zunächst in einer Datenbank verschiedene Formulare unabhängig voneinander erfasst und später zusammengemischt werden. Dann müssen aber die PNr und der zusätzliche Code von Anfang an sorgfältig in alle Bögen eingetragen werden, gewissermaßen als doppelte Buchführung.
In doppelt blinden Studien stellt die Klinikumsapotheke anhand einer Randomliste die Präparate her und beschriftet sie nur mit dem Studiennamen und einer laufenden Nummer. Diese Nummer muss in jedem Fall als Patienten-ID benutzt werden, damit später bei der Auswertung eine Zuordnung möglich wird.

6.1.3. Identifikatoren für das Formular

Sie dienen dazu, einen Fragebogen oder ein anderes Formular eindeutig zu identifizieren. Meistens benutzt man dafür einen Code aus Ziffern und/oder Buchstaben mit fester Länge (nicht zu lang, 3-5 Zeichen reichen). Manche Datenbanksysteme benötigen diesen Schlüssel, um Daten aus verschiedenen Formularen korrekt zusammenzufügen (zu mischen). Wird das gleiche Formular mehrfach zu verschiedenen Zeitpunkten verwendet (bestimmte Labordaten werden während der Studie 3mal erhoben), so vergeben Sie für jeden Zeitpunkt eine eigene ID (z.B. Lab1, Lab2 und Lab3). In der EDV bedeuten verschiedene Formulare in der Regel verschiedene Dateien bzw. Datenbanktabellen mit einem eigenen Namen. Für die Auswertung müssen meistens Daten aus mehreren Tabellen zusammengefügt werden. Dann sollte sich der Name des Formulars auch im Variablennamen niederschlagen (s.u.).

6.1.4. Identifikatoren für Items (Variablen)

Natürlich muss auch jedes Item (d.h. jeder Wert bzw. jede Antwort) eindeutig identifizierbar sein. Dies ist oft die schwierigste Aufgabe, weil - je nach verwendetem Programm - die dafür erlaubte Zeichenzahl sowie die erlaubten Zeichen stark eingeschränkt sind. Praktisch kein Programm erlaubt die Verwendung von Leerzeichen, diverse Sonderzeichen dürfen nicht verwendet werden (z.B. Rechenzeichen, oft auch nationale Sonderzeichen wie ö oder à), oder bestimmte Worte sind tabu (z.B. AND oder NOT). Die Namen müssen immer eindeutig sein, auch wenn Items aus verschiedenen Datensätzen zusammengemischt werden. So könnten z.B. in einer Erhebung „Geburtshilfe“ sowohl im Fragebogen „Mutter“ als auch im Bogen „Kind“ Merkmale wie Geburtsdatum, Gewicht oder Größe vorkommen. Wenn Sie in beiden Bögen als Identifikator (Variablenname) GEBDAT verwenden, ergeben sich Konflikte (und Fehlermeldungen), wenn Sie versuchen, die Daten von Mutter und Kind in einem Datensatz zu vereinen. In solchen Fällen müssten Sie von vornherein unterschiedliche Namen vergeben, beispielsweise MGEB und KGEB. Am besten benutzen Sie grundsätzlich die ersten Stellen des Variablennamens, um den Fragebogen zu identifizieren, das erleichtert später auch die Fehlersuche, da Sie sofort erkennen, aus welchem Bogen bzw. welcher Datei die Items stammen. Gleiche Inhalte sollten immer gleich benannt werden (ergänzt um das Kürzel für den Fragebogen). Falls später 1:n - Relationen geplant sind (z.B. verschiedene Zeitpunkte nebeneinander), müssen die verschiedenen Wiederholungen ebenfalls durch den Namen zu unterscheiden sein (z.B. L03Chol

für Laborbogen, 3. Messung, Cholesterin). Angenommen, in einer Tumorstudie soll bei der Aufnahme, nach 3 und 6 Zyklen Chemotherapie sowie bei Therapieabschluss eine Kernspintomographie durchgeführt werden. Das Ergebnis der Untersuchung nennen Sie dann z.B. AufNmrE, Ch3NmrE, Ch6NmrE und AbsNmrE. Die ersten 3 Zeichen identifizieren den Zeitpunkt, der Rest den Inhalt. Für das Datum der Untersuchung verwenden Sie die gleichen Namen und ersetzten das E am Ende durch ein D. Natürlich können Sie auch andere Namen wählen oder die Kennung für den Fragebogen ans Ende stellen. Es ist aber bestimmt nicht sinnvoll, die Variable erst AufBildgDiag, dann Chemo3MagResResult und vielleicht noch Zyk6MRT zu nennen, damit macht man sich die Auswertung selbst unnötig schwer. Bei Studien mit sehr vielen Items müssen Sie möglicherweise so viele Informationen in den Variablennamen packen (Wiederholungen), dass kaum noch ein sinnvoller Name vergeben werden kann. Manche verzichten dann ganz auf mnemotechnische Bezeichnungen und verwenden stattdessen Bezeichnungen wie z.B. A0234 oder D0001 (Fragebogen A, Frage 234 bzw. Fragebogen D, Frage 1). Nachteil: Wird hier eine Variable nachträglich eingefügt, müssen auch viele Variablennamen geändert werden.

6.2. Standardisierung

Standardisierte Bögen zur Datenerfassung erleichtern die Einhaltung des Datenplans und des zeitlichen Protokolls. Ein guter Fragebogen kann entscheidend sein für den Erfolg einer Studie. Das gilt natürlich in besonderem Maße, wenn die Befragten die Bögen selbst ausfüllen sollen. Weder das Überfordern („ich verstehe überhaupt nicht, was die von mir wollen“) noch das Unterfordern („wollen die mich mit dem Bogen veralbern?“) der Interviewten trägt dazu bei, dass diese die Fragebögen sorgfältig ausfüllen.
Standardisierung innerhalb der Studie bedeutet, dass alle erfassten Werte vergleichbar sein müssen. Das gilt sowohl für verschiedene Orte und Personen (multizentrische Studie, nicht immer der gleiche Arzt untersucht den Patienten) als auch für verschiedene Zeitpunkte (Messwiederholungen, Nachuntersuchungen). Dass die Dimension von Messwerten immer die gleiche sein muss, versteht sich wohl von selbst. Aber auch die Verwendung von verschiedenen Messgeräten, die Einführung neuer Untersuchungsmethoden etc. können zu unerwünschten Effekten führen (es sei denn, sie sind die Zielgröße der Studie).
Standardisierung bedeutet aber auch das Erscheinungsbild innerhalb des Fragebogens sowie von verschiedenen Bögen. Bei Multiple Choice Fragen z.B. ist es Geschmackssache, ob Sie Kästchen oder Kreise ankreuzen lassen oder ob die Befragten die Codes direkt in ein Zahlenfeld schreiben. Das einmal gewählte Format sollte aber immer beibehalten werden, wie auch die Felder zum Eintragen von Zahlen immer gleich aussehen sollten. Werden für eine Studie mehrere Fragebögen verwendet, gilt das einheitliche Design für alle Bögen. Außerdem sollten sie ein einheitliches Deckblatt haben, das einerseits die Zugehörigkeit zur Studie und andererseits den Typ des Bogens sofort erkennen lässt. Nach Möglichkeit sind auch unterschiedliche Papierformate zu vermeiden. Übergroße Tabellen (maximal A3) erschweren die Arbeit unnötig.
Die Zahl der unterschiedlichen Formate sollte nicht größer sein als unbedingt nötig. Es ist sehr wichtig, dass die Probanden sich schnell im Fragebogen zurechtfinden.
Besonders in multizentrischen oder gar multinationalen Studien müssen die Dimensionen streng festgelegt und deren Einhaltung überwacht werden. Gewicht in Pfund oder Kilogramm? Oder sind gar anglophile Partner beteiligt, die noch in lbs (englischen Pfund) messen? Labordaten in internationalen Einheiten, in mg oder ml? Sind alle Laborgeräte übereinstimmend geeicht? Besonders gefährlich wird es natürlich, wenn die Werte aufgrund verschiedener Maßeinheiten nur geringfügig differieren, da resultierende Fehler dann leicht als Effekte ausgelegt werden können. Falls die Partner sich jeweils weigern, ein anderes Maß als bisher üblich zu verwenden, müssen die Daten umgerechnet werden, bevor sie in einer Datei zusammengefügt werden können. Dabei müssen Sie natürlich die Übersicht behalten, damit die gleichen Daten nicht versehentlich mehrfach umgerechnet und somit wieder falsch werden.

6.3. Codierung der Werte

Als erstes ist zu definieren, welche Items benötigt werden und welche Antworten oder Ausprägungen jeweils möglich sind. Dabei spielt die Fragestellung eine wichtige Rolle. Es müssen genaue Regeln für die Codierung aufgestellt werden, um die Einheitlichkeit zu gewährleisten.

6.3.1. Quantitative Daten

Quantitative Daten sind gemessene oder gezählte Größen, die als Zahlenwert erfasst werden, z.B. Körpergröße, Blutdruck, Alter oder die Anzahl der Infekte in einem Jahr. Da es sich um echte Zahlen handelt, kann mit ihnen gerechnet werden, d.h. Sie können geeignete numerische Verfahren anwenden wie Berechnung von Mittelwert und Standardabweichung oder Varianzanalysen. Welche Verfahren das im Einzelnen sind, richtet sich nach dem Zusammenhang und der Art der Daten (gezählt oder gemessen, schief oder normalverteilt, ...).
Geben Sie die Werte so ein, wie Sie sie erhoben haben. Je genauer die Messungen sind, umso besser (aber nicht übertreiben!). Klassen bilden Sie bei Bedarf später im Auswertprogramm. Achten Sie - besonders bei der Verwendung von Datenbanken - darauf, dass die Werte auch wirklich als Zahlen und nicht im

Textformat eingegeben werden. Und selbstverständlich müssen alle Werte die gleiche Dimension haben. Dies gilt nicht nur für alle Fälle innerhalb einer Datenspalte, sondern auch bei Messwiederholungen. Wenn Sie bei zweijährigen Kindern die Körpergröße in cm messen und drei Jahre später in Metern, werden Sie eine drastische und statistisch signifikante Schrumpfung bei allen Kindern feststellen.

6.3.2. Zeit und Datum

Das Hantieren mit Datum und/oder Uhrzeit ist meistens recht umständlich. Wenn Sie nur herausbekommen möchten, zu welcher Stunde die meisten Kinder geboren werden, ist das kein Problem. Oft benötigen Sie aber die Differenz zwischen zwei Zeitangaben, z.B. zur Berechnung des Alters, einer Überlebenszeit, der Liegedauer auf einer Station oder der Dauer einer Geburt. Wenn die Wehen am 31.12 beginnen und das Kind am 1.1. geboren wird, müssen Sie dazu alle Zeitangaben (Minute, Stunde, Tag, Monat und Jahr) berücksichtigen. Eine zunehmende Zahl von Programmen erleichtern Ihnen die Arbeit, indem sie ein Zeit/Datum-Format anbieten. Zunächst müssen Sie für die betreffende Variable den Typ „Datum“ bzw. „Zeit“ auswählen. In den meisten Programmen können Sie anschließend festlegen, in welchem Format das Datum bzw. die Uhrzeit angezeigt wird (mit Wochentag, Monat als Zahl, abgekürzt oder ausgeschrieben, Jahr zwei- oder vierstellig etc.). In diesem Format geben Sie das Datum auch ein, das Programm wandelt Ihre Eingabe selbständig in das interne Speicherformat um. Falsche Angaben, etwa den 29. 2. 2005 oder den 35. Mai wird das Programm zurückweisen. Achten Sie aber unbedingt auf die Reihenfolge von Tag und (numerischem) Monat! 4/6/2005 könnte sowohl der 4. Juni als auch der 6. April sein! Das Jahr sollten Sie immer vierstellig eingeben, da verschiedene Programme zweistellige Zahlen unterschiedlich ergänzen.

6.3.3. Qualitative Daten

Qualitative Daten enthalten eine Aussage, nicht quantitativ verwertbare Zahlen, d.h. Sie können mit ihnen nicht wirklich rechnen. Das gilt auch dann, wenn Sie für diese Variablen wie empfohlen numerische Codes verwenden. Es ist leicht einzusehen, dass die Bildung eines mittleren Geschlechts (codiert mit 1 für männlich und 2 für weiblich) wenig sinnvoll erscheint. Die Möglichkeiten für qualitative Variablen sind beschränkt (Auszählung, Kontingenztafel), oft dienen sie auch als Gruppenvariable, um den Datensatz in mehrere unverbundene Stichproben aufzuteilen. Schließlich müssen Sie noch unterscheiden, ob Sie es mit nominalen (Aufzählung ohne Ordnung, z.B. rot, grün, blau, gelb) oder mit ordinalen (Folge, z.B. sehr gut, gut, mäßig, schlecht) Daten zu tun haben.

Nehmen wir an, dass der Anfang eines Fragebogens folgendermaßen aussieht:

Fragebogen für das Projekt XYZ

Code-Nummer: _____

Kind geboren am _____

Geschlecht _____

(usw.)

Mögliches Ergebnis der Auszählung für das Merkmal „Geschlecht“:

Ausprägung	Anzahl	Interne Codierung (ASCII)
Bruder	2	066 114 117 100 101 114
Fräulein	1	070 114 132 117 108 101 105 110
Junge	2	074 117 110 103 101
M	3	077
M	1	077 032
Mädchen	2	077 132 100 099 104 101 110
Männlich	8	077 132 110 110 108 105 099 104
Schwester	1	083 099 104 119 101 115 116 101 114
W	7	087
Weiblich	6	087 101 105 098 108 105 099 104
m	6	109
maennlich	2	109 097 101 110 108 105 099 104
männl	1	109 132 110 110 108
männl.	3	109 132 110 110 108 046
männlich	15	109 132 110 110 108 105 099 104

w	9	119
weibl.	5	119 101 105 098 108 046
weiblich	19	119 101 105 098 108 105 099 104
Summe	93	

Eigentlich kann ein Computer keinen Text verarbeiten, geschweige denn verstehen. Texte werden gespeichert, indem die Tasten auf einer Schreibmaschine durchnumeriert und die entsprechenden Zahlen im Speicher abgelegt werden. Bei der Ausgabe „drückt" dann der Computer die entsprechenden Tasten in der richtigen Reihenfolge, so dass der gewünschte Text auf dem Papier erscheint. Dabei bekommt nicht nur jeder Großbuchstabe, jeder Kleinbuchstabe, jede Ziffer und jedes Satzzeichen eine Codezahl, sondern u.a. auch der „Wagenrücklauf" (Carriage return) und - besonders wichtig - die Leertaste (Blank).
Inhaltlich kann ein Rechner mit den Zahlen nichts anfangen. Ihnen mag es egal sein, ob das Wort „männlich" groß oder klein geschrieben ist, der PC erkennt nur zwei nicht identische (also unterschiedliche) Folgen von Zeichencodes (*Zeichenketten*, englisch *Strings*, genauer *character strings*), da ein m den Code 109 bekommt, ein M dagegen 77. Die Strings „männl" und „männl." unterscheiden sich durch den Punkt (Code 46) am Ende sogar in der Länge (5 bzw. 6 Zeichen). Und dass in der Liste scheinbar zwei verschiedene Ms auftauchen liegt daran, dass dreimal nur ein M eingetippt wurde (Code 77), einmal aber ein M und zusätzlich noch ein Leerzeichen (Codes 77 und 32).
In jedem Fragebogen gibt es Items, die keine gemessene Zahl liefern, sondern verbal beantwortet werden. Manche lassen vielleicht nur die Antworten „ja" und „nein" zu, bei anderen erzählt Ihnen jeder Proband etwas anderes (z.B. „Welche Krankheiten hatten Sie bisher?"). In dem Beispiel oben (Geschlecht des Kindes) haben wir für zwei Geschlechter 19 verschiedene Antworten bekommen, wobei sich beliebig viele weitere Varianten inklusive Tippfehler ausdenken lassen. Und steht der Buchstabe M eigentlich als Abkürzung für „Männlich" oder für „Mädchen"?
Angenommen, Sie möchten untersuchen, ob Kneippsche Anwendungen bei Kindern eine positive Auswirkung auf das Immunsystem haben. Wenn Sie die Eltern nach früheren Erkrankungen ihrer Kinder fragen, erzählen sie Ihnen vielleicht von den Fußballverletzungen, erwähnen aber nicht die häufigen Erkältungen, die sie im Winter als völlig normal ansehen. Sie müssen also unbedingt bei der Planung festlegen, welche Punkte zur Klärung der Fragestellung wichtig sind und dann gezielt danach fragen (aber nicht suggestiv!).
Generell gilt: wenn alle möglichen Antworten auf eine Frage bekannt sind, sollten diese im Fragebogen aufgeführt und codiert werden. Dies ist natürlich nur für eine überschaubare Zahl von Antworten praktikabel, nicht bei 50 oder gar 200. Solche Items sind aber normalerweise ohnehin nicht sinnvoll, da sie nicht statistisch auswertbar sind, sondern nur zu Einzelfallbeschreibungen führen. Im Extremfall haben Sie für einzelne Fragen mehr mögliche Antworten als Probanden in der Studie. Zu viele oder auch zu lange Antworten (ganze Sätze oder gar mehrere Absätze) machen sogar eine tabellarische Auflistung der Ergebnisse nicht sinnvoll.
Am besten ist es, wenn Sie eine Liste möglicher Antworten vorgeben, aus denen die Befragten eine aussuchen. Dabei ist darauf zu achten, dass diese immer vollständig und disjunkt sind.

- Die Fragen und die Antworten müssen so formuliert sein, dass die Befragten sie auch verstehen. Formulieren Sie so kurz wie möglich und so ausführlich wie nötig.
- Die Antworten müssen immer disjunkt sein, d.h. es darf keine Überschneidungen bzw. Mehrfachantworten geben. Es ist technisch unmöglich, mehrere Angaben in einer Variablen zu speichern. Teilen Sie die Frage ggf. auf.
- Die Antwortmöglichkeiten müssen vollständig sein, d.h. alle Befragten müssen eine passende Antwort im Angebot finden. Wenn die Probanden die Frage nicht verstehen bzw. keine passende Antwort finden, fragen sie entweder nach (Zeitaufwand), lassen die Frage unbeantwortet (Informationsverlust) oder sie kreuzen irgend etwas an (Informationsverfälschung). Notfalls kann man auf dem Bogen Raum für zusätzliche Bemerkungen lassen und dann selbst entscheiden, welche Antwort zutrifft. Es ist aber selten sinnvoll, solche Anmerkungen im Klartext in den Computer einzugeben.
- Trennen Sie andererseits aber auch keine inhaltlich zusammengehörenden Antworten auf mehrere Fragen auf.
- Verwenden Sie für ähnliche Fragen immer die gleichen Codes (z.B. 0 für nein und 1 für ja). Es ist üblich, Verneinungen (nein, nicht, nie) mit 0 zu verschlüsseln. In anderen Fragen (Geschlecht, Augenfarbe) beginnt man die Codierung mit 1. Codieren Sie die Antworten, nicht die Inhalte. Ein „nein" bekommt z.B. immer den Code 0, egal ob die Frage „fühlen Sie sich wohl" oder „hatten Sie Schmerzen" lautet.

In manchen Fällen existieren bereits fertige Schlüssel. Diese sind oft sehr detailliert und hierarchisch aufgebaut (z.B. Diagnosenschlüssel nach ICD) und für Studien mit einer konkreten Fragestellung meistens zu kompliziert. Sie können aber als Vorlage gute Dienste leisten.
Am Beispiel oben haben Sie gesehen, dass ein Rechner nicht in der Lage ist, eingegebenen Text zu verstehen und Synonyme oder Tippfehler zu erkennen. Der zweckmäßigste Weg ist nach wie vor, die Antworten einfach durchzunummerieren und dann diese Zahlen einzugeben. Die Codierung als Zahl wird

zwar von manchen als rückständig empfunden, aber ein Computer kann mit Zahlen sehr viel schneller und effektiver umgehen als mit Textzeichen. Vergessen Sie aber nie, dass es sich trotzdem um eine nominale bzw. ordinale Größe handelt und Sie folglich keine statistischen Verfahren anwenden dürfen, die numerische Messwerte erwarten.
Viele Gründe sprechen für eine numerische Codierung:

- Wie schon gesagt, Computer können mit Zahlen viel besser und vor allem schneller umgehen als mit Text (Zeichenketten). Um z.B. alle Patienten mit dem Geschlecht „männlich" herauszusuchen, muss für jeden einzelnen Fall der gesamte Text Zeichen für Zeichen mit der Vorgabe verglichen werden. In diesem Beispiel sind also pro Fall 8 Vergleichsoperationen nötig, bei numerischer Codierung nur eine einzige.
- Bei Verwendung von Klartext müssen gleiche Antworten absolut identisch sein. Jede abweichende Schreibweise (Synonyme, Abkürzungen, Großschreibung, Leerzeichen, Tippfehler) gilt als neue Ausprägung der Variablen (s.o.).
- Bei den üblichen Statistikpaketen ist die Verwendung von Zeichenketten oft stark eingeschränkt. Auszählungen oder Kreuztabellen von Stringvariablen sind meistens möglich, aber wenn Sie Gruppenvergleiche anstellen wollen, erhalten Sie früher oder später die Fehlermeldung „Strings sind hier nicht erlaubt" und müssen doch zur numerischen Codierung übergehen.
- Kein Programm kann erkennen, dass „gut" besser ist als „mangelhaft". Wenn Sie alle Prüflinge zusammenfassen möchten, die bestanden haben, müssten Sie als Bedingung ***IF Note = „sehr gut" OR Note = „gut" OR Note = „befriedigend" OR Note = „ausreichend"*** formulieren. Bei der üblichen Codierung von Schulnoten geht das einfacher mit ***IF Note < 5***. Das spart nicht nur Tipparbeit, sondern bei der Ausführung auch Rechenzeit ein. Tests für ordinale Werte wie Cramers Phi können Sie nur mit numerisch codierten Variablen berechnen.

Damit die erzeugten Ergebnisse besser lesbar sind, können Sie bei vielen Programmen zusätzlich die Bedeutung der Codes eingeben. Diese meistens als ***Labels*** oder ***Etiketten*** bezeichneten Texte werden aber nicht in der Datenmatrix, sondern zusammen mit den Variablennamen im (meistens nicht direkt lesbaren) Kopf der Datei gespeichert. In den Daten selbst steht also in der Spalte „geschl" eine 1 oder eine 2. Irgendwo merkt sich das Programm aber, dass „geschl" für „Geschlecht", die 1 für „männlich" und die 2 für „weiblich" steht. In den Ergebnissen werden dann beim Geschlecht die Zahlen 1 und 2 automatisch durch den entsprechenden Text ersetzt oder ergänzt.

Fragen sollten immer so formuliert werden, dass eine Antwort zutrifft. Wenn Sie nur ein Feld 'ja' zum Ankreuzen vorsehen, wissen Sie bei einem leeren Kästchen nicht, ob die Antwort definitiv 'nein' lautet, ob der Proband die Frage übersehen hat oder sie nicht beantworten wollte; d.h. die Unterscheidung von 'nein' und '*fehlend*' ist nicht möglich.

Beispiel 1:

Wie oft haben Sie die Therapie in den letzten 4 Wochen durchgeführt?

- ☐ (5) täglich
- ☐ (4) fast täglich
- ☐ (3) etwa jeden 2. Tag
- ☐ (2) 1-2mal pro Woche
- ☐ (1) noch seltener, aber mindestens einmal
- ☐ (0) niemals
- ☐ (9) keine Angabe

Die 6 möglichen Antworten schließen einander aus, und alle Befragten sollten eine für sie zutreffende Antwort in der Liste finden.

Beispiel 2:

Welche Personen leben im Haushalt?

- ☐ (1) Mutter
- ☐ (2) Vater
- ☐ (3) Kind(er)
- ☐ (4) Großeltern
- ☐ (5) sonstige
- ☐ (9) keine Angabe

Die Liste der Antworten ist vollständig, die Kategorie „sonstige" deckt alle weiteren Möglichkeiten ab. Die Antworten sind aber **nicht disjunkt**. Bei vielen Familien müssten die Codes 1, 2 und 3 gleichzeitig in eine Zelle geschrieben werden, was technisch unmöglich ist. Da alle 5 Angaben gleichzeitig auftreten können, müssen Sie 5 Datenspalten (Variablen) einrichten, in die jeweils der Antwortcode 1 für „ja" oder 0 für „nein" eingetragen wird. Außerdem sollte im Fragebogen immer eine Antwort gefordert werden, sonst wissen Sie später nicht, ob die Befragten mit „nein" geantwortet oder die Fragen übersehen haben:

Welche Personen leben im Haushalt?

	ja (1)	nein (0)	unbekannt (9)
Mutter	☐	☐	☐
Vater	☐	☐	☐
Kind(er)	☐	☐	☐
Großvater/ -mutter	☐	☐	☐
weitere Personen	☐	☐	☐

Beispiel 3:

Röntgenbefund der Lunge

	ja (1)	nein (0)	unbekannt (9)
ohne Befund	☐	☐	☐
links	☐	☐	☐
rechts	☐	☐	☐
beidseitig	☐	☐	☐

Hier wurde eine Aussage in vier Fragen (d.h. vier Datenspalten) aufgeteilt. Das ist sehr verwirrend, erschwert die Auswertung und führt zu Fehlern. Eine solche Konstruktion fordert unsinnige Antworten geradezu heraus. Was ist, wenn jemand bei allen Punkten „ja“ ankreuzt? Aber auch gutwillige Menschen werden verwirrt. Muss bei beidseitigen Auffälligkeiten nur „beidseitig“ oder auch „links“ und „rechts“ mit ja beantwortet werden? Sinnvoll wäre eine einzige Frage mit vier möglichen Antworten:

Röntgenbefund der Lunge

- ☐ (0) ohne Befund
- ☐ (1) nur links auffällig
- ☐ (2) nur rechts auffällig
- ☐ (3) beidseitig auffällig
- ☐ (8) Befund unklar
- ☐ (9) keine Untersuchung

Mit einer Bedingung wie ***IF (lunge = 1 OR lunge = 2)*** könnten jetzt alle Patienten mit einseitigem Befund zusammengefasst werden.

Beispiel 4:

<u>Ein- und Ausschlusskriterien:</u>

Gesunde Probanden im Alter von 20 bis 35 Jahren	☐
Herzfrequenz unter 55/Minute	☐
Schwangerschaft oder Stillzeit	☐
Teilnahme an einer anderen klinischen Studie	☐
Klinisch relevante oder instabile Erkrankungen	☐

Diese Liste ist äußerst problematisch und kann zur Ablehnung der ganzen Studie führen! Denn wann soll welches Kästchen angekreuzt werden? Wenn ein Kreuz ‚ja‘ bedeutet, heißt ein leeres Kästchen bei den letzten drei Punkten nicht automatisch ‚nein‘, vielleicht hat der Arzt nicht danach gefragt. Dies wäre aber ein krasser Verstoß gegen das Studienprotokoll (z.B. Aufnahme einer schwangeren Frau). Man könnte alle Fragen so formulieren, dass die Bedingung immer erfüllt sein muss (*keine* Schwangerschaft, *keine* Teilnahme an einer anderen Studie). Besser ist es in jedem Fall, Antwortfelder für ‚ja‘ und ‚nein‘ vorzusehen, dann besteht kein Zweifel mehr, ob eine Frau schwanger ist oder nicht, bzw. ob der Punkt noch zu klären ist.

<u>Ein- und Ausschlusskriterien:</u>

Gesunde Probanden im Alter von 20 bis 35 Jahren	⓪ nein	① ja
Herzfrequenz unter 55/Minute	⓪ nein	① ja
Schwangerschaft oder Stillzeit	⓪ nein	① ja
Teilnahme an einer anderen klinischen Studie	⓪ nein	① ja
Klinisch relevante oder instabile Erkrankungen	⓪ nein	① ja

6.3.4. Scores und Skalen

Häufig ergibt sich das Problem, Dinge zu erfassen, die eigentlich nicht erfassbar sind. Typisch in der Medizin sind hier u.a. Skalen, die das psychische Befinden der Patienten (Lebensqualität) messen sollen und die in der Regel von qualifizierten Gremien in langen Sitzungen erarbeitet, ausgiebig getestet und schließlich freigegeben (validiert) werden. Trotzdem werden Sie keine Skala finden, an der es nichts zu kritisieren gibt. Es ist aber zweifelhaft, ob Sie deshalb unbedingt eine neue entwickeln müssen, die höchstwahrscheinlich auch nicht besser wird. Die Verwendung von validierten Fragebögen ermöglicht außerdem, die Ergebnisse der eigenen mit anderen Studien zu vergleichen (mit der nötigen Vorsicht!). Beachten Sie, dass für die meisten Skalen ein Copyright besteht und vor der Verwendung in eigenen Studien eine Genehmigung eingeholt und eventuell eine Gebühr bezahlt werden muss.
In der Medizin wird oft versucht, aus mehren Einzelfaktoren, die Einfluss auf ein bestimmtes Ereignis haben, eine einzige Maßzahl ***(Score)*** zu errechnen. Ein typisches Beispiel wäre das APGAR-Schema, bei dem für ein Neugeborenes aus den Punkten **A**tmung, **P**uls, **G**rundtonus, **A**ussehen und **R**eflexe eine Zahl zwischen 0 (leblos) und 10 (optimal vital) bestimmt wird. Nicht immer ist die Formel für den Score so einfach. Oft müssen Faktoren entsprechend ihrer Bedeutung unterschiedlich gewichtet werden. Die meisten Statistikpakete können neue Variablen berechnen, Sie können also die einzelnen Faktoren eingeben und später die Scores vom Programm ausrechnen lassen. Voraussetzung dafür ist, dass Sie die Formel kennen und sie nicht zu kompliziert ist; rekursive Berechnungen sind mit Programmen wie SPSS nicht möglich.

6.3.5. Nachträgliches Codieren

Nicht immer wird es möglich sein, die Codierung direkt bei der Datenerhebung vorzunehmen. Dann müssen zunächst die Originalbelege gesammelt und vor der Dateneingabe die benötigten Informationen herausgesucht und verschlüsselt werden. Zu bedenken ist, dass bei einer Interpretation durch Fremde (also nicht die direkt Befragten) immer zusätzliche subjektive Faktoren dazukommen. Nach Möglichkeit sollte die Codierung nicht durch die Initiatoren der Studie erfolgen. Sonst ist die Gefahr sehr groß, dass z.B. ein Arzt in den Röntgenbildern nur die Dinge erkennt, die er sehen möchte, auch wenn er sich um Objektivität bemüht, zumal wenn er die Patienten auch noch selbst behandelt.
Aus einer größeren Zahl von Fragebögen werden die Antworten im Klartext notiert und von einem Gremium klassifiziert, d.h. sie suchen nach typischen Antworten, fassen diese zusammen und bilden daraus eine verhältnismäßig kleine Zahl von möglichen Antworten. Anhand dieser Liste werden dann alle Fragebögen codiert. Das erfordert natürlich einen weitaus höheren personellen Aufwand und erhöht die Gefahr, dass die Codierer die Antworten falsch interpretieren. Falls weiterhin Daten für die Studie erhoben werden, sollten die Fragebögen entsprechend geändert werden. Wenn bei einigen Items die Antwort „sonstige“ mit dem Zusatz „wenn ja, welche? _____“ besonders häufig ausgefüllt wurde, haben Sie vielleicht wichtige Antworten vergessen und ein manuelles Nachcodieren ist notwendig.
Eine andere Möglichkeit ist, dass es sich um Bilder oder Tonaufnahmen handelt. In diesem Fall ist eine Codierung durch eine entsprechende Fachkraft unerlässlich. Ein Laie wird kaum in der Lage sein, ein Röntgenbild, ein Ultraschalldiagramm oder das Bild eines Kernspintomografen zu klassifizieren. Der Arzt muss aus dem Bild die für die Studie relevanten Faktoren entnehmen und so objektiv wie möglich den entsprechenden Items und Codes zuordnen. Das Verdichten der Daten ist eine verantwortungsvolle Aufgabe, die eine Maschine normalerweise nicht vollbringen kann. Damit ein Computer eine solche Arbeit sinnvoll erledigen kann, braucht er ein meistens sehr umfangreiches Regelwerk, nach dem er vorgehen kann. Die Entwicklung solcher Programme mit allen Tests und Probeläufen kann Jahre dauern. Ein Bild ist für einen Rechner nichts weiter als ein Wust aus Millionen von Pixeln (Punkten) unterschiedlicher Helligkeit und Farbe. Schon das Erkennen einfacher Strukturen (Linien, Texte, Kreise etc.) erfordert viel Aufwand. Auch noch relevante von unwichtiger Information zu unterscheiden, ist für eine Rechenmaschine kaum möglich. Es ist nicht praktikabel, solche Programme individuell für eine kleine Studie anzufertigen.
Wenn mehrere Personen - vielleicht auch noch an verschiedenen Orten - Daten codieren, ist eine regelmäßige Absprache unumgänglich, um systematische Fehler zu vermeiden.
Falls normalerweise geschultes Hilfspersonal die Codierung vornimmt, in Zweifelsfällen aber Experten gefragt werden müssen, kann dafür ein spezieller Code eingetragen werden. Dieser wird einerseits im Auswertprogramm als fehlender Wert deklariert, erinnert aber auch daran, diese Fälle noch dem Experten vorzulegen.

6.4. Fehlende Werte

Fehlende Daten müssen so codiert werden, dass jede Verwechslung mit gültigen Werten ausgeschlossen ist. Gerade in medizinischen Arbeiten sind fehlende Angaben kaum zu vermeiden. Ob der Patient bestimmte Aussagen verweigert, Pannen im Labor passieren, einzelne Untersuchungen aufgrund äußerer Umstände nicht möglich sind oder ob Probanden einer Langzeitstudie in Urlaub waren, es gibt zahlreichen Gründe, warum einzelne Angaben fehlen können. Spätestens das Auswertprogramm muss diese von gültigen Aussagen unterscheiden können, um gravierende Fehler zu vermeiden. Fast alle Statistikprogramme und

auch manche Datenbanken kennen für numerische Werte einen speziellen Code für fehlende Angaben (*SPSS* bezeichnet ihn z.B. als *System Missing Value* oder kurz ***SYSMIS***, die Datenbank *Access* als ***Null***). Eingegeben oder angezeigt wird er üblicherweise als leeres Feld, als isoliertes Dezimalzeichen (Punkt bzw. Komma ohne eine Ziffer) oder durch das entsprechende Schlüsselwort (z.B. *SYSMIS* oder *NULL*). Dieser Leerwert wird in Statistikprogrammen auch allen Zellen zugewiesen, die ungültige Werte erhalten (falsches Format beim Einlesen der Daten) oder Ergebnisfeldern, wenn die in der Formel benutzten Operanden fehlende Werte enthalten oder das Ergebnis ungültig ist, etwa bei Division durch 0. Wenn die Datenerfassung nicht direkt im Statistikprogramm erfolgt, muss nicht nur das benutzte Eingabeprogramm, sondern auch alle eventuell dazwischengeschaltete Software mit Missing Values umgehen können.
Beispiel: Sie erfassen Ihre Daten mit *Microsoft Access*, das den Leerwert *Null* kennt, die Auswertung soll mit *SPSS* erfolgen. Sie könnten in *Access* eine Datei im Format *dBase III* erzeugen und diese wieder in *SPSS* einlesen. Da dBase III aber keine *Missings* kennt, werden aus allen „*Null*"-Werten in Access (als Wort geschrieben) numerische Nullen (als Zahl). Kommt in den Items die Null auch als gültiger Wert vor, ist eine Unterscheidung von „0" und „keine Angabe" nicht mehr möglich! In diesem Beispiel könnten Sie als bessere Alternative eine Tabelle im *Excel*-Format herausschreiben (vorausgesetzt die Zahl der Variablen ist nicht zu groß) oder Sie benutzen *ODBC* (s. Kapitel über Datenbanken).
Eine Alternative zu den automatischen Leerwerten ist es, fehlende Werte selbst mit Codes zu versehen. Bei Messwerten muss natürlich eine Schlüsselzahl verwendet werden, die nie als Wert auftreten kann. Noch aus der Zeit der Lochkarten stammt die Benutzung der Null (z.B. bei Körpergröße oder Gewicht) oder das Auffüllen der vorgesehenen Spaltenbreite mit Neunern (z.B. 99 für Zahl der Kinder oder 999 als Blutdruck). Falls in der ganzen Studie nirgends negative Werte auftauchen können, bieten sie eine gute Möglichkeit zur Codierung von Missings. *Vorteil:* Das oben beschriebene Problem bei der Umwandlung von Dateiformaten entfällt. Außerdem ist es of erforderlich, mehrere Gründe für das Fehlen einer Angabe zu unterscheiden (z.B. -1 = wurde nicht erhoben, -2 = entfällt, -3 = Angabe verweigert, -10 = ungültiger Wert im Fragebogen). *Nachteil*: Bei der Eingabe ist etwas mehr Tipparbeit erforderlich und im Auswertprogramm müssen die fehlenden Werte explizit deklariert werden, damit sie nicht als gültige Werte in die Berechnungen (z.B. von Mittelwerten) einbezogen werden. Da ist es natürlich einfacher, wenn alle Werte kleiner Null *fehlend* sind, als wenn für jedes Item andere Codes definiert wurden. Einige dieser Fehlwerte könnten bei Verwendung einer Datenbank mit Eingabemaske automatisch eingesetzt werden, z.B. -1, wenn nichts ausgefüllt wurde, -2 wenn Felder aufgrund bestimmter Antworten automatisch übersprungen wurden und -10 wenn der eingetippte Wert außerhalb des Gültigkeitsbereichs liegt.

6.5. Richtlinien für das Formulardesign

Es ist durchaus sinnvoll, die Daten direkt auf dem Originalbogen computergerecht zu codieren. Die Daten für den Computer sollten auffällig sein, damit beim Erfassen nicht der ganze Bogen gelesen werden muss. Bewährt haben sich genügend große Kästchen, die übersichtlich postiert werden.
Ein ***Item*** ist ein Formularelement, das zur Aufnahme einer einzelnen Antwort dient. Es besteht aus erklärendem Text und einem Feld für die Antwort. Der Text kann aus der eigentlichen Frage, den möglichen Antworten und Erläuterungen dazu bestehen.

- Die Antwortfelder sollten so angelegt sein, dass sie sowohl beim Ausfüllen des Bogens als auch beim späteren Eingeben der Daten auffallen und leicht zu finden sind. Sie müssen nicht unbedingt am äußeren Ende des Bogens liegen (bei großen Abständen zwischen Frage und Antwortfeld kann das zu Fehlern führen), sollten aber deutlich umrandet und bündig untereinander angeordnet sein.
- Benutzen Sie keine Schrift unter 10 Punkt.
- Das Mindestmaß für Kästchen für auszufüllende Antwortfelder ist 0,5*0,7cm, Ankreuzfelder sollten deutlich kleiner sein.
- Trennen Sie die einzelnen Items optisch voneinander. Lassen Sie also mehr Platz zwischen 2 Items als zwischen den einzelnen Punkten eines Items. Die inhaltliche Gliederung des Bogens soll sich auch im Erscheinungsbild darstellen.
- Es kann sinnvoll sein, Überschriften, die möglichen Antworten und zusätzliche Erläuterungen durch den Druck hervorzuheben (verschiedene Zeichensätze, -größen oder -attribute), z.B. Überschriften fett, die Antworten normal und Erläuterungen kursiv. Aber nicht übertreiben und einheitlich für alle Bögen!
- Wenn Items in Abhängigkeit von gegebenen Antworten übersprungen werden, sollte das jeweilige „Ziel" leicht zu finden sein (Pfeile, besonders hervorgehobene Items, Rahmen etc.).
- Formulieren Sie den Text so knapp und präzise wie möglich. Stichwortartige Fragen werden manchmal besser verstanden als lange, grammatikalisch korrekte Sätze.
- Passen Sie Ihre Sprache der Klientel an. Für Mediziner müssen Sie Ihre Fragen anders formulieren als für Kinder in der Grundschule. Benutzen Sie gängige Vokabeln. Bei der Befragung von Rauschgiftabhängigen werden Sie mit „Druck" und „Turkey" eher Antworten erhalten als mit „intravenösem Betäubungsmittelabusus" und „Entzugssyndrom".

- Der Fragebogen sollte Personen testweise vorgelegt werden, die in ihrem Wissensstand der zu befragenden Klientel entspricht. Sonst kann es leicht passieren, dass alle Ärzte oder Psychologen den Bogen für gut befinden, ein „Normalsterblicher“ aber praktisch kein Wort versteht.
- Wenn nötig, sehen Sie Felder für „unbekannt“ oder „nicht zutreffend“ vor. Vermeiden Sie es, dass die Probanden bei einer Frage nichts ankreuzen. Ausnahme: Die fehlenden Antworten ergeben sich aus dem Kontext („Wenn Sie Frage 3 mit ‘nein’ beantwortet haben, bitte weiter bei Frage 12“).
- Geben Sie bei Zahlenfeldern die erwartete Stellenzahl als Kästchen und bei Dezimalzahlen die Position des Kommas vor. So erreichen sie bei allen Messungen die gleiche Genauigkeit.
- Bei Datumsangaben geben Sie das gewünschte Format vor (TMJ oder MTJ?). Lassen Sie Monate ggf. ausschreiben, um Missverständnisse Tag/Monat zu vermeiden.
- Wenn eine Rubrik „sonstige“ vorgesehen ist, lassen Sie Platz für eine kurze Antwort im Klartext. Es ist in der Regel nicht sinnvoll, diese statistisch auszuwerten. Es kann aber sein, dass eine oder mehrere Antworten gehäuft auftreten. Dann müssten Sie die Codierung dieses Items erweitern und die Daten im Rechner entsprechend ergänzen.
- Für Antworten im Klartext geben Sie einen Kasten oder eine Linie vor, aber keine einzelnen Kästchen für jeden Buchstaben. Letzteres erschwert nicht nur das Schreiben, sondern auch das Lesen der Information. Das kann natürlich dazu führen, dass die Antwort länger ausfällt als das dafür vorgesehene Textfeld in der Datenbank. Nach Möglichkeit sollten Sie auf solche Freitextfelder bei der Auswertung ohnehin verzichten, da sie schwer auszuwerten sind (Groß/Kleinschreibung, Synonyme, Tippfehler, Abkürzungen).
- Bedenken Sie, dass es für fehlende Angaben manchmal mehrere Möglichkeiten gibt (vergessen, verweigert, nicht zutreffend, nicht durchgeführt, weiß nicht). Es kann sinnvoll sein, zwischen diesen zu unterscheiden.
- Testläufe sind immer empfehlenswert. Sie dienen nicht nur der Überprüfung des Fragebogens, sondern auch dem Training der Interviewer.

Wenn Sie einen Fragebogen fertig gestellt haben, müssen Sie Ihr schlimmster Kritiker sein. Stellen Sie sich vor, Sie haben absolut keine Lust auf Fragebögen und füllen Sie dann Ihr eigenes Machwerk aus. Welche Fragen können Sie nicht oder falsch verstehen? Vermeiden Sie Zweckoptimismus („wird schon irgendwie gut gehen“), sondern denken Sie an Murphy’s Gesetz!

7. Einsatz von EDV

7.1. Allgemeine Überlegungen

Die Benutzer von Computern verbringen oft viele Stunden mit der Lösung von Problemen, die sie ohne Computer gar nicht hätten.

PCs oder sonstige Computer sind **Maschinen!** Sie erledigen manches erheblich schneller und exakter als der Mensch und führen die gleiche langweilige Aufgabe tausendmal hintereinander aus, ohne dabei zu ermüden. Kein Mensch kann in einer Sekunde eine Million Berechnungen ausführen, dafür kann ein Rechner aber auch eine Million Fehler pro Sekunde fabrizieren! Wenn vorher fähige Programmierer gute Arbeit geleistet haben, kann ein Computer rekursive Berechnungen in Minuten erledigen, an denen ein Mensch Monate sitzen würde. Andererseits erkennt selbst ein Laie oft auf den ersten Blick eine Fraktur auf einem Röntgenbild, an dem auch ein gutes Bilderkennungsprogramm lange herumrechnen würde. Ein Programm, das Texte fließend und fehlerfrei von einer Sprache in eine andere übersetzt, ohne dabei den Sinn zu entstellen, gibt es bis heute noch nicht. Denn eines kann der PC mit Sicherheit nicht: Denken. Glauben Sie nicht, dass der Computer ein Wundertier ist, das alle Ihre Fehler automatisch korrigiert. Halten Sie ihn vielmehr für das, was er wirklich ist: einen willigen Trottel, der alle Ihre Befehle ohne jeden Widerspruch wortwörtlich ausführt - ohne Rücksicht darauf, ob sie sinnvoll sind oder gar die eigene Arbeit vernichten.

Die Auswertung einer Studie per Computer erfordert also doppelte Sorgfalt. Nur wenn die Studie inhaltlich gut geplant ist**,** die erhobenen Daten computergerecht erfasst werden und die verwendete Software fehlerfrei und passend ist, können Sie mit brauchbaren Ergebnissen rechnen. Normalerweise muss für die Aufbereitung der Daten und die Bereinigung von Fehlern (soweit überhaupt möglich) deutlich mehr Zeit und Geld aufgewendet werden als für die eigentliche Auswertung. Und inhaltliche Fehler kann auch der teuerste Supercomputer nicht korrigieren.

Jeder Einsatz von Computern kostet zunächst einmal Zeit und Geld, auch wenn die Preise für PCs in den letzten Jahren bei gleichzeitiger Leistungssteigerung ständig gefallen sind. Die Kosten für die benötigte Software kann den Wert der Hardware schnell übersteigen; vor allem Spezialprogramme, die nur in kleiner Stückzahl vertrieben werden, sind recht kostspielig.

Dazu kommt dann der oft erhebliche Zeitaufwand, um sich in die Programme einzuarbeiten. Auch (oder gerade) die auf den ersten Blick so benutzerfreundlichen *Windows*-Programme erfordern eine hohe Einarbeitungszeit, besonders wenn dem Anwender die üblichen Standardfunktionen nicht genügen. Da aus falscher Sparsamkeit meist keine gedruckte Dokumentation mehr mitgeliefert wird, muss sich auch der ehrliche Käufer die benötigten Informationen mühsam aus der sogenannten Online-Hilfe heraussuchen. Und das Schreiben und anschließende Testen von individuellen Programmen - auch in Form von Makros und/oder Visual BASIC etc. - ist auch für erfahrene Computerbenutzer zeitaufwendig.

Schließlich sind Programme Werkzeuge und ersetzen nicht nötiges Fachwissen. Ein gutes Notensatzprogramm macht den Benutzer nicht automatisch zu Beethoven, ein CAD-Programm macht keinen Architekten und Statistiksoftware keinen Biometriker.

Der EDV-Einsatz sollte also gründlich überlegt werden. Natürlich können Sie für eine Handvoll Daten erst eine Datenbank und eine Eingabemaske erstellen, die Datei dann in das Format eines Statistikprogramms konvertieren und mit diesem schließlich einen Chi-Quadrat-Test rechnen. Das bedeutet aber, dass Sie möglicherweise etliche Stunden vor dem Rechner verbringen, um dann bei der Auswertung 10 Minuten zu sparen. Ein bisschen Handarbeit um die Zellenhäufigkeiten auszuzählen und ein Taschenrechner für den Test wären ausreichend gewesen. Die Bewältigung von großen Datenmengen oder rechenaufwendige Verfahren wie z.B. logistische Regression sind andererseits ohne EDV nicht möglich. Der zusätzliche Aufwand, den der Einsatz von Computern mit sich bringt, sollte immer in einem sinnvollen Verhältnis zur dadurch eingesparten Zeit und Arbeit stehen.

Noch einmal: je besser eine Studie geplant und vor allem je klarer die Fragestellung formuliert ist, umso einfacher wird die Auswertung sein. Wer irgendwelche Daten sammelt und diese irgendwie irgendwo eingibt, um möglichst viele Signifikanzen zu berechnen, darf auch vom Computer keine sinnvolle Hilfe erwarten. Das wäre etwa das Gleiche, wie wenn jemand einen Ziegelstein aufs Gaspedal legt, sich im Sitz zurücklehnt und sein Auto auffordert, für eine erholsame Urlaubsreise zu sorgen.

7.2. Datenformate

Intern besteht ein Computerspeicher aus Zellen - den sogenannten ***Bytes*** - die jeweils eine Zahl zwischen 0 und 255 aufnehmen können. Alle Informationen wie z.B. größere Zahlen oder solche mit Nachkommastellen, Texte, Töne, Grafiken und Programme müssen als eine entsprechende Zahlenfolge codiert werden.

Dabei unterscheidet man folgende Standardtypen:

1. ***Integer:*** Speichert nur ganze Zahlen, aber mit Vorzeichen. Mit einem Byte kann man also die Zahlen von -128 bis +127 darstellen. Ein einfacher Integerwert fasst 2 Bytes zu einer Zahl zusammen und umfasst so einen Wertebereich von -32768 bis +32767. 4 Bytes (32 Bits) ergeben zusammengefasst einen Wert vom Typ ***Long Integer*** im Bereich von mehr als ± 4 Milliarden. Zahlen, die den erlaubten Wertebereich

sprengen (z.B. Multiplikationsergebnis), führen zu einem Prozessorfehler und im schlimmsten Fall zum Absturz des Computers, wenn das Programm den Fehler nicht sauber abfängt (ein solcher Integer-Überlauf hat schon zum Absturz einer *Ariane*-Rakete geführt!). Besonders kritisch ist die Division, da Nachkommastellen einfach entfallen (z.B. 11:6 = 1).

2. ***Realzahlen*** können im Gegensatz zu Integerwerten Nachkommastellen haben. Eine solche Zahl speichert der Computer intern in zwei Teilen: Die Ziffernfolge ab der ersten Ziffer ungleich Null und die Position des Dezimalpunkts. Die Zahl 0.00000057352 würde als .57352E-06 abgelegt, der Punkt muss um 6 Stellen nach links verschoben werden. Das E wie Exponent steht als Abkürzung für „mal 10 hoch ...“: $0.57352 * 10^{-6} = 0.00000057352$. Wie viele Ziffern gespeichert werden ist festgelegt, weitere werden einfach weggerundet. Bei Messwerten ist das meistens kein Problem, Sie werden kaum Ihre Urlaubsreise auf den cm genau planen oder erwachsene Menschen auf ein Mikrogramm genau wiegen. Die im PC benutzten Prozessoren (seit i486DX) können solche Zahlen direkt verarbeiten (‘Fließkommaarithmetik’, ***FPU***). Durch die Rundung kann es abhängig von der Rechengenauigkeit problematisch werden, wenn sehr kleine und sehr große Zahlen miteinander verrechnet werden.
3. ***Datum und Zeit:*** Eine zunehmende Zahl von Programmen erleichtern den Umgang mit Datums- und Zeitangaben: Sie geben das Datum und/oder die Uhrzeit in einem bestimmten Format (z.B. 10. Dezember 1995 14:32 oder auch 10. 12. 95 14:32) ein und bekommen sie auch so angezeigt. Intern benutzen die Programme aber ein spezielles Format, meistens in Form einer Zeitdifferenz zu einem fixen Stichtag. Welcher Tag und welches Format das ist, unterscheidet sich von Programm zu Programm. Wenn es eine Möglichkeit gibt, Daten zwischen Programmen auszutauschen, werden aber fast immer auch diese Angaben korrekt umgewandelt. Das Format *„Date“* (nur Datum), *„Time“* (nur Uhrzeit) oder *„DateTime“* (Datum und Uhrzeit) müssen Sie in den meisten Programmen ausdrücklich einstellen.
4. ***Text:*** Ein Computer kann Text weder lesen noch verstehen. Alle verfügbaren Textzeichen (Groß- und Kleinbuchstaben, Ziffern, Satzzeichen, Leerzeichen) sowie ein paar wichtige Funktionen (Wagenrücklauf, neue Zeile) bekommen einen Zahlencode zwischen 0 und 255 zugewiesen. Jeder zusammenhängende Text ist für einen Computer nur eine Aneinanderreihung solcher Textcodes, also eine ***Zeichenkette*** oder englisch ein ***character string*** oder kurz ein ***string***.
5. ***Logische oder boolesche Felder*** kennen nur die beiden Möglichkeiten ja (*wahr*, *true*) oder nein (*falsch*, *false*). Nein wird üblicherweise mit 0 codiert, ja mit 1. In grafischen Oberflächen erscheinen sie oft als Kästchen mit oder ohne Haken. Da boolesche Variablen keine weiteren Zustände kennen (fehlende Werte, ‚weiß nicht‘), ist meist die Benutzung eines Integerfeldes vorzuziehen.

Wichtig: Numerische Daten speichert ein Computer so, dass er sie ohne Umwandlung sofort verarbeiten kann, d.h. im internen, binären Zahlenformat des Prozessors. Für einen Menschen wäre es recht mühsam, die bis zu 64stelligen 0/1-Folgen richtig zu interpretieren. Daten werden bei der Ausgabe auf Bildschirm oder Drucker automatisch in ein lesbares dezimales Format umgewandelt. Umgekehrt geben Sie Zahlen über die Tastatur als Text ein und das Programm wandelt diesen in die entsprechende interne Darstellung (*Integer*, *Real* oder *Date*) um. Fehlerhafte Eingaben werden vom Programm sofort zurückgewiesen. Wenn Sie Zahlenfelder im Eingabeprogramm als Textfeld definieren, werden Sie auf dem Bildschirm kaum einen Unterschied bemerken (Text erscheint meistens linksbündig, Zahlen dagegen rechtsbündig). Wenn Sie später aber mit den Werten rechnen wollen, ist das unmöglich, weil sie vom Prozessor nicht verarbeitet werden können. Viele Programme sind nicht in der Lage, die Umwandlung nachträglich durchzuführen. Und wenn doch, kann es zu Datenverlusten kommen, weil fehlerhafte Eingaben (z.B. unerlaubtes Datum, Dezimalpunkt statt -komma) oft einfach zu fehlenden Werten gemacht werden.

8. Datenerfassung

8.1. Der Datenfluss

Der Begriff Datenfluss umfasst den gesamten Weg der Daten von der Versuchsplanung über das Auswertprogramm bis hin zu den Ergebnissen. Dazu gehören u.a. die Fragen

- Welche Daten werden in welcher Form gebraucht?
- Wie kommen die Daten in den Computer? Welche Geräte und Medien werden dafür benötigt? Wer gibt die Daten ein, wo und womit (PC, Programme).
- Werden die Daten ein zweites Mal eingegeben oder wie werden sie sonst überprüft?
- Was geschieht mit fehlerhaften oder zweifelhaften Daten? Wer ist zuständig bei Rückfragen? Werden „Stellvertreter" eingegeben oder der ganze Datensatz bis zur Klärung zurückgestellt?
- Sollen erfolgte Korrekturen dokumentiert werden und wenn wie?
- Reicht kommerzielle (fertige) Software, oder müssen individuelle Programme entwickelt werden?
- Wie wird das Eingabepersonal geschult?

Vor allem bei umfangreichen Studien mit verschiedenen Fragebögen und Datensätzen ist die Planung und die Dokumentation des Datenflusses sehr wichtig. Sonst ist die Gefahr groß, dass Sie sich irgendwann in Ihren eigenen Daten nicht mehr auskennen. Welche der vielen Dateien enthält die korrigierten Daten in der aktuellen Fassung? Nicht immer ist garantiert, dass das die Datei mit dem jüngsten Datum ist. Und gibt es noch Fragebögen, die neu dazugekommen bzw. korrigiert, aber noch nicht in den Computer eingegeben wurden?

8.2. Datenqualität

Neben der Kontrolle des Datenflusses ist es natürlich genauso wichtig, dass die Daten selbst korrekt erhoben, auf den Computer übertragen und ausgewertet werden. Die Möglichkeiten, hierbei Fehler zu machen, sind nahezu unerschöpflich:

- Abweichungen vom vorgeschriebenen Protokoll oder Missverständnisse.
- Nicht geeignete oder nicht geeichte Messinstrumente.
- Ungenaue, unleserliche oder unvollständige Aufzeichnung der Daten.
- Fehlerhafte oder unvollständige Erfassung auf elektromagnetischen Datenträgern.
- Fehlerhafte Zuordnung von separat oder nachträglich erhobenen Daten (falsche Patientennummer), Verwechslung von Laborbögen.
- Gefährdung der Datenqualität durch exzessive Datensammlungen.
- Ungenügend ausgebildetes oder geschultes Personal.
- Nachträgliche nicht oder nicht vollständig dokumentierte Änderungen.
- Fehler beim Benutzen oder Programmieren von Datenbanken.
- Missbrauch statistischer Software, Gebrauch von nicht angemessenen statistischen Verfahren.
- Verlust von Daten oder Programmen durch Beschädigung oder Verlust der Datenträger (Brand, Headcrash, beim Transport beschädigte Medien, Softwarefehler, Computerviren).

Wenn für eine Studie beliebig viel Zeit und Geld zur Verfügung stehen, kann der Idealzustand einer absolut fehlerfreien Datei annähernd erreicht werden (jeder Laborwert wird mehrfach bestimmt, alle aufgeschriebenen oder eingegebenen Daten werden mehrfach überprüft etc.). In der Praxis wird es kaum Datensammlungen ohne Fehler geben. Eine gute Planung des Datenflusses kann die Fehlerrate deutlich senken. Bei der Prüfung beachtet man vor allem die Daten, die besonders relevant für die Studie sind. In einer randomisierten Studie ist das z.B. die Durchführung der Randomisierung und die Zuordnung der Probanden zu den Gruppen. Für wichtige Daten gilt der Leitsatz „prüfe frühzeitig und häufig".
Eine hundertprozentige Fehlererkennung ist praktisch nicht möglich. Falsche Werte, die formal korrekt und plausibel sind, können grundsätzlich nie erkannt werden (außer durch Vergleich mit den Originaldaten). Manche Fehler können Sie aber durch formale und logische Überprüfung entdecken, bevor die eigentliche Auswertung beginnt. Besonders elegant geschieht dies mit eigens für jeden Fragebogen entwickelten Eingabemasken eines Datenbanksystems. Das Erstellen der Masken ist aber zeitaufwendig und lohnt sich kaum bei einfach strukturierten und übersichtlichen Daten (s. Abschnitt über Datenbanken). Als Alternative können auch die Funktionen eines Statistikpakets dienen. Dabei sollten Sie auf keinen Fall nur die Windowsoberfläche benutzen, sondern die nötigen Befehle als Kommandodatei abspeichern. Nur dann können Sie die Checks nach der Datenkorrektur oder beim Hinzukommen neuer Fälle in unveränderter Form erneut durchführen und bei Bedarf mit neuen Befehlen ergänzen. Die gängigsten Prüfungen sind:

- ***Korrekter Datentyp**:* Wird bei der Eingabe der Werte automatisch überprüft. Vorausgesetzt, Sie stellen von Anfang an den richtigen Datentyp (numerisch, Text, Datum) ein.
- ***Wert liegt innerhalb des möglichen Wertebereichs:*** Alle wichtigen Variablen auszählen lassen oder wenigstens Minimum und Maximum bestimmen und nachschauen, ob unmögliche Werte dabei sind. In SPSS bietet sich für quantitative Zahlen die Funktion *EXAMINE* an.

- ***Konsistenz von Items innerhalb eines Fragebogens:*** Für qualitative Variablen Kreuztabellen, für quantitative Punktwolken erstellen lassen und nach auffälligen Werten suchen. Oder beispielsweise nur die Männer selektieren und alle gynäkologischen Variablen auszählen. Erscheinen nur fehlende Werte oder gibt es etwa auch gültige Angaben?
- ***Konsistenz von Items in verschiedenen Fragebögen:*** Es gilt das Gleiche wie oben, vorher müssen die Daten aber in eine Datei zusammengeführt werden.
- ***Konsistenz von Items über die Zeit (bei wiederholten Erhebungen):*** Hier könnte man zusätzlich die Differenzen bilden und auf plausible Werte überprüfen.
- ***Korrekte Probanden-ID:*** Zusätzlich zur Fallnummer (die zum Zusammenführen von Dateien ausschließlich benutzt wird) können in einer anderen Variablen zusätzliche Informationen gespeichert werden, z.B. Geburtstag und/oder Initialen. Diese müssen dann für alle Fälle mit der gleichen Nummer identisch sein.
- Ausführliche Datenchecks (alle Bögen vollständig vorhanden, keine Widersprüche bei doppelt vorhandenen Angaben wie Geschlecht, Alter usw.) sind aufwendig bei der Programmierung und auch bei der Ausführung. Um eine zügige Dateneingabe zu gewährleisten, müssen solche Checks meist als separate Jobs ausgeführt werden, entweder in regelmäßigen Abständen oder aus besonderen Anlässen (z.B. vor der Auswertung oder Weitergabe der Daten).

8.3. Methoden und Werkzeuge zur Datenerfassung

Unter Datenerfassung versteht man die Umwandlung der erhobenen Daten in eine Form, die von Maschinen gelesen, gespeichert und verarbeitet werden kann. Meistens heißt das, dass die auf Papier geschriebenen Daten mit Hilfe der Tastatur und eines geeigneten Programms eingetippt werden. Das kann direkt das für die Auswertung vorgesehene Statistikprogramm sein. Häufig wird für die Eingabe aber ein anderes Programm verwendet, entweder weil dieses mehr Komfort bietet oder weil das Auswertprogramm zum Zeitpunkt der Datenerfassung nicht zur Verfügung steht. Dann ist aber Vorsicht geboten: Erstens muss es eine Möglichkeit geben, die Daten ohne jede Verfälschung (Missing Values, Datumsvariablen) in das Format des Auswertprogramms zu überführen. Zweitens müssen Sie sich immer an der rechteckigen Datenmatrix orientieren. Es macht wenig Sinn, die Daten in einer Form zu erfassen, mit der SPSS oder ein anderes Statistikpaket nachher nichts anfangen kann.
Der Ablauf der Datenerfassung sollte gut organisiert sein. Die erfolgte Eingabe ist auf den Bögen zu vermerken. Die eingegebenen Belege werden zunächst gesammelt. Nachdem von den Daten erfolgreich eine Sicherungskopie erstellt wurde, ordnen Sie die Fragebögen in die vorgesehenen Aktenordner ein. (Sollte die aktuelle Datei einmal zerstört werden, laden Sie die letzte Sicherung und geben die losen Bögen noch einmal ein. Das ist bedeutend einfacher, als alle abgehefteten Bögen zu überprüfen).
Wenn sich die Studie über einen längeren Zeitraum erstreckt, beginnen Sie mit der Datenerfassung, während noch Daten erhoben werden. Dadurch wird nicht nur die Zeit bis zur Auswertung verkürzt, regelmäßige Backups verringern die Gefahr eines Verlustes. Fehler im Fragebogen werden rascher erkannt und können ggf. korrigiert werden.

Die wichtigsten Möglichkeiten für die Datenerfassung sind:

8.3.1. Eintippen der Daten direkt in das Auswertprogramm

Das ist die einfachste Methode, vorausgesetzt das Programm verfügt über einen brauchbaren Dateneditor. In *SPSS* können Sie die Variablen definieren (Name, Typ und Etiketten) und anschließend die Werte direkt in die Datenmatrix eintippen. *Vorteil:* Sie brauchen weder ein zusätzliches Programm zu beschaffen noch den Umgang damit zu erlernen. Korrekturen erfolgen ebenfalls direkt in der Datenmatrix. Es gibt nur eine gültige Version Ihrer Daten und aufwendige und fehleranfällige Umwandlungen von Dateiformaten entfallen. *Nachteil:* Keine individuell gestalteten Eingabemasken, keine Überprüfung der Werte bei der Eingabe (außer Datentyp).

8.3.2. Datenbanksysteme

Vorteil: Individuell gestaltete Eingabemasken, diverse Prüfungen direkt bei der Eingabe, verschiedene Fragebögen werden als selbständige Tabellen erfasst und erst später zusammengeführt. Das relationale Modell erlaubt bei der Datenerfassung die Abweichung von der Rechteckmatrix. *Nachteil:* Höherer Aufwand, mögliche Probleme beim Umwandeln des Dateiformats. Das relationale Modell wird schnell zum Nachteil, wenn zunächst viele Daten ohne klares Konzept gesammelt werden und eine sinnvolle Auswertung anschließend nicht möglich ist. Mehr zum Thema Datenbanken folgt in einem Extrakapitel.

8.3.3. Tabellenkalkulationen und andere Software

Die Benutzung von Drittsoftware bringt in der Praxis kaum inhaltliche Vorteile. Diesen Weg können Sie gehen, wenn Sie Ihre Daten zu Hause eingeben wollen, aber SPSS nicht besitzen. Anstatt die Datenmatrix direkt in SPSS einzutippen, können Sie z.B. auch ein Tabellenkalkulationsprogramm verwenden, wenn es Dateien im *Excel*-Format anlegen kann. Das Arbeitsblatt muss im Prinzip genauso aussehen wie die SPSS-

Datenmatrix. Formatierungen wie Fettschrift oder Farbe gehen grundsätzlich verloren. Andererseits muss das Zellenformat dem Datentyp entsprechen; Zahlen in Textform werden auch als String eingelesen und ein Datum wird nur korrekt importiert, wenn die Zelle auch als Datum formatiert ist. Falls Sie von den Fähigkeiten des Kalkulationsprogramms Gebrauch machen, um neue Variablen zu berechnen, müssen Sie vor dem Export der Daten alle versteckten Formeln durch die Ergebnisse ersetzen. Alle Variablendefinitionen (Name, Format, Labels) gehen verloren und müssen neu eingegeben werden.

8.3.4. Daten als Rohtext (ASCII)

Diese Form der Dateneingabe stammt aus der Lochkartenzeit, trotzdem wird sie manchmal noch benutzt. Die Daten werden als reiner Text (ASCII oder ANSI codiert, ohne jede Formatierung) in Tabellenform gespeichert. Hier müssen also nicht nur die gleichen Items immer untereinander stehen, sondern auch jeweils die Einer-, Zehner-, Hunderterstelle usw., z.B. ID in Spalte 1-4, Geschlecht in Spalte 5, Alter in Spalte 6-7. Bevor Sie festlegen, welcher Wert wo zu stehen hat (*Schlüsselliste*), müssen Sie also überlegen, wie viele Stellen jedes Item maximal belegen kann. Dezimalzeichen lassen Sie grundsätzlich weg, erst beim Einlesen der Daten legen Sie fest, ob die Ziffernfolge „183“ als 183 cm oder als 1,83 m einzulesen ist. Bei vielen Daten pro Fall dürfen die Zeilen umgebrochen werden, die Anzahl der Datenzeilen (Records) pro Fall muss aber immer gleich sein. Beschränken Sie sich beim Zeichensatz unbedingt auf den genormten Teil des ASCII-Codes (7 Bits, Code 0 bis 127), das sind die Ziffern und das englische Alphabet. *Vorteil:* Dieses Format kann praktisch jedes Statistikprogramm lesen. Eine solche Datei braucht nur wenig Platz auf der Platte und kann auch zwischen verschiedenen Plattformen (Großrechner, Mac, PC) übertragen oder uncodiert über das Internet verschickt werden. *Nachteil:* Sowohl die Eingabe der Daten über einen Editor als auch das Einlesen durch ein Programm sind umständlich, alle Variablendefinitionen (Name, Format, Labels) gehen verloren und müssen neu eingegeben werden.

8.3.5. Datenerfassungsbüros

In professionellen Datenerfassungsbüros arbeiten oft Teilzeitkräfte, die ein paar Euro nebenbei verdienen wollen und nach Anschlägen bezahlt werden. Ihnen ist es egal, ob sie Bilanzen, Kundenbestellungen oder Daten einer medizinischen Studie eintippen. Wichtig ist vielmehr, dass sie die einzugebenden Werte deutlich erkennen können und sie nicht mühsam im Fragebogen zusammensuchen müssen. Hier ist es also besonders wichtig, dass die Werte deutlich geschrieben und im Fragebogen klar gekennzeichnet sind und nach Möglichkeit bündig untereinander stehen. Erfassungsbüros benutzen in der Regel eigene Programme, Schnelligkeit ist hier wichtiger als eine optisch schöne Maske. Auch Plausibilitätsprüfungen sind nicht sehr sinnvoll, da das Personal ohnehin nicht weiß, wie es auf falsche Werte im Fragebogen reagieren soll. Geübte Datentypisten geben die Daten mit hoher Geschwindigkeit blind ein, d.h. sie schauen nur auf den Fragebogen und weder auf die Tastatur noch auf den Bildschirm. Dabei sind Tippfehler selten, aber natürlich nicht ganz auszuschließen. Statt nun mühsam und zeitaufwendig die Werte auf dem Bildschirm und auf dem Fragebogen optisch zu vergleichen, werden (auf Wunsch und gegen Aufpreis) alle Daten ein zweites Mal eingetippt (möglichst von einer anderen Person), mit den bereits vorhandenen Daten verglichen und Diskrepanzen angezeigt. Dieses Verfahren verhindert aber nur Tippfehler, falsch ausgefüllte Felder oder unleserlich geschriebene Zahlen führen nach wie vor zu Fehlern. Plausibilitätsprüfungen müssen in jedem Fall später von Mitarbeitern der Studie vorgenommen werden.

8.3.6. Computer assisted data collection (CADC)

Die Daten gelangen - ohne Umweg über Papier etc. - direkt in den Computer, z.B. über Internetmasken oder durch die Interviewer mit einem Notebook bei einer Telefonbefragung bzw. direkt vor Ort. Dabei ist es natürlich wichtig, wie die Daten eingegeben werden und von wem:

- Die häufigste Methode ist die, dass die Interviewer die Antworten nicht auf Papier notieren, sondern direkt in ein Notebook (z.B. mit Hilfe einer Datenbankmaske) eingeben.
- Erheblich mehr Aufwand ist erforderlich, wenn die Befragten den Computer selbst bedienen sollen. Einerseits müssen die Rechner sicher vor Manipulationen durch Computerfreaks, andererseits auch von Benutzern ohne Computerkenntnissen sicher und einfach zu bedienen sein. Dazu ist oft spezielle Hardware wie z.B. Touchscreens erforderlich, ähnlich wie beim Geldautomaten einer Bank.

Datenerhebung und -eingabe werden zu einem Schritt zusammengefasst.

Vorteile:

- Ein Arbeitsschritt (vom Papier zum Computer) entfällt. Das spart Zeit und Geld; außerdem entfällt eine potentielle Fehlerquelle.
- Sofort durchgeführte Plausibilitätsprüfungen ermöglichen sofortige Rückfragen.
- Automatisches Springen beim Ausfüllen.
- Berechnungen von neuen Größen sind sofort möglich (ggf. mit entsprechender Reaktion).

Nachteile:
- Die von den Probanden, von Prüfärzten etc. unterschriebenen Bögen sind wichtige Dokumente, bei denen nachträgliche Korrekturen - im Gegensatz zu Veränderungen in einer Datei - meistens nachweisbar sind. In vielen Studien sind schriftliche Unterlagen daher vorgeschrieben.
- Es passiert viel leichter, dass jemand aus Versehen die Nachbartaste erwischt (z.B. 5 statt 2 im Ziffernblock), als dass jemand eine '5' auf den Bogen schreibt, wenn die '2' gemeint ist. Ein Nachschlagen in Originalbelegen ist nicht möglich!
- Bei Papiervorlagen können nachträgliche Korrekturen deutlich dokumentiert werden (so durchstreichen, dass der alte Wert noch erkennbar ist, neuen Wert daneben schreiben und quittieren). Außerdem ist es möglich, bei Zweifeln Randbemerkungen anzufügen.
- Datenverluste bei einem Hardwaredefekt, Systemabsturz, durch bösartige Viren oder gar Diebstahl des Notebooks sind immer ärgerlich. Bei herkömmlicher Datenerhebung existieren die Daten in verschiedenen Formen (z.B. Papier, Rohdaten, auf USB-Sticks usw.) und können notfalls noch einmal neu erfasst werden. Elektronisch erfasste Daten sind unrettbar verloren, wenn keine externen Kopien existieren. Häufige Backups werden für die Studie lebenswichtig!
- Nicht immer ist CADC praktikabel. Eine MTA wird z.B. nicht ständig zum PC laufen, um jedes Ergebnis einzeln einzutippen, sondern die Werte immer auf ein Blatt Papier schreiben.

8.3.7. Maschinenlesbare Belege

Eine andere Möglichkeit ist der Einsatz von maschinenlesbaren Belegen, besonders bei Fragebögen mit Multiple-Choice-Fragen. Die Bögen müssen dann aber sorgfältig ausgefüllt und behandelt werden, damit sie nicht durch Flecken, Knicke etc. optisch nicht mehr gelesen oder mechanisch nicht mehr eingezogen und transportiert werden können. Zur Erfassung von Zahlen sind sie weniger geeignet, da hier die häufigsten Fehler auftauchen. Das Lesen von ausgefüllten Bögen mit Hilfe eines Scanners und üblichen Texterkennungsprogrammen (OCR Software) ist trotz stark verbesserter Programme immer noch fehleranfällig und langsam. Es gibt Software, mit denen Fragebögen erstellt und – wenn sie ausgefüllt wurden – eingescannt und ausgelesen werden können. Diese Programme sind nicht billig, und für einen sinnvollen Einsatz müssen Sie auch in die Hardware investieren, insbesondere in schnelle Scanner mit zuverlässigem Vorlageneinzug.

8.3.8. Automatische Erfassung

Manche Laborgeräte können die Ergebnisse direkt auf einen Datenträger schreiben. Das ist natürlich effizienter und weniger fehleranfällig als das Abschreiben der Werte auf Papier. Zu bedenken ist jedoch, dass manche Geräte eine wahre Datenflut liefern, die dann schwer zu verarbeiten ist (z.B. EKG über 24 Stunden) und aus der die für die Studie relevanten Werte extrahiert werden müssen. Außerdem erlauben die Geräte nicht immer die Eingabe eines Identifizierungscodes für den Patienten, sondern vergeben eigene Nummern. Dann kann es mühsam und fehleranfällig sein, die Daten den richtigen Patienten zuzuordnen. Oft werden auch die Daten verschiedener Untersuchungen auf einen Datenträger geschrieben; der Arzt muss dann wissen (und notieren), dass der 19. Wert zum Patienten Peter Müller gehört. Werden für eine Erhebung mehrere solcher Maschinen eingesetzt, so müssen also die Daten für einen Patienten aus mehreren Datenträgern zusammengesucht werden. Schließlich muss es eine physikalische Möglichkeit geben, die Daten auf einen PC zu übertragen (Kabel, Netzwerk, Datenträger). Das Format, in dem der Automat die Daten erfasst und speichert, ist selten PC-kompatibel, was zusätzliche Hard- und Software nötig machen kann (Interfaces, Adapter, Treiber, Umwandlungsprogramme).

8.4. Umwandlung von Datenformaten

Nahezu jedes Programm benutzt beim Speichern seiner Daten einen spezifischen Code, das heißt im Prinzip kann ein Programm nur die Dateien einlesen, die es zuvor selbst erstellt hat. Das gilt oft auch für verschiedene Versionen des gleichen Programms, da ständig neue Funktionen dazukommen, die irgendwie auch im Dateiformat (d.h. bei der Codierung der Daten) berücksichtigt werden müssen. In der Regel ist das Format abwärtskompatibel.
Viele Programme können aber einige fremde Dateiformate lesen (***importieren***) oder auch schreiben (***exportieren***), das heißt sie werden vollständig in das Format des anderen Programms umgewandelt. Für gängige Formate gibt es oft Lösungen, die allgemein verwendet werden und keinem spezifischen Programm zugeordnet sind (sog. *Meta-Dateien*). Typische Beispiele sind das *Rich-Text-Format .rtf* für Texte oder Grafikformate wie *.jpg, .tif* oder *.gif. Vorteil:* Die Daten können uneingeschränkt mit den Werkzeugen des neuen Programms bearbeitet werden. *Nachteil:* Alles, womit das neue Programm nichts anfangen kann, geht verloren. Wenn Sie beispielsweise in Word eine Tabelle erstellt haben und in ein Textprogramm ohne Tabellenfunktionen importieren, so finden Sie in diesem allenfalls den Inhalt (also den eigentlichen Text) der Tabelle wieder. Mit Sicherheit werden Sie aber an den Formatierungen einiges nachbessern müssen. Bei Daten für statistische Auswertungen klappt die Übertragung der Werte meistens problemlos, falls es die entsprechenden Import/Exportfilter gibt. Etwas schwieriger wird es bei Datums- und Zeitangaben. Und nur

selten gelingt die Übertragung aller Angaben im Dateikopf, d.h. Variablennamen, *Labels* für Variablen und deren Ausprägungen und die Definition von *Missing Values*. Eine Variante ist ***ODBC***, bei der sich z.B. ein Statistikprogramm die Daten direkt aus einer Datenbank holt.

8.5. Umgang mit fehlerhaften Daten

Mit der Entdeckung fehlerhafter oder unvollständiger Daten ist es nicht getan. Besonders wenn an der Studie mehrere Personen oder gar Zentren beteiligt sind, muss auch der Umgang mit fehlerhaften Daten genau geplant und festgelegt werden (*Error Reports, Data Queries*).

- Was geschieht mit nur teilweise ausgefüllten Items? Besonders häufig werden Datumsangaben aus der Erinnerung nur unvollständig gegeben („das war irgendwann im Juli 1990“). Gilt dann das ganze Feld als *missing* oder wird einfach ein (immer gleicher) Tag eingesetzt? Entscheidend ist dabei, wie groß der Fehler im Verhältnis zur Aussage werden kann. Bei der Berechnung der Liegedauer wäre es nicht akzeptabel, einen fehlenden Aufnahmetag durch eine „1“ zu ersetzten. Wenn der Patient am 28. entlassen wird, ergäbe sich so eine Liegedauer von 28 Tagen, auch wenn der Patient wirklich nur 3 Tage auf der Station war. Handelt es sich bei dem beobachteten Zeitraum dagegen um Jahre, ist eine Ungenauigkeit von 15 Tagen nicht relevant.
- Wichtig ist auch, dass bei der Eingabe nicht einfach aus Bequemlichkeit gültige Werte eingegeben werden. Steht auf dem Fragebogen beispielsweise eine 3, die Eingabemaske erlaubt aber nur die Codes 1 und 2, so wird manches Mal einfach irgend ein gültiger Wert eingetippt.
- Es ist wichtig, dass Sätze mit zweifelhaften Daten nicht in die Auswertung gelangen. Entweder, die Daten kommen erst in die endgültige Datenbank, wenn alle Eingaben überprüft und fehlerfrei sind, oder für alle Datensätze wird ein „Valid“-Feld eingefügt, das nur bei einwandfreien Daten gesetzt wird. Dann kann einerseits die Auswertung nur auf die validen Daten beschränkt werden, andererseits gezielt nach zu korrigierenden Datensätzen gesucht werden. Verwendet man kein logisches, sondern ein numerisches Feld, so können mehrere Fehlerzustände dokumentiert werden (0=alles ok, 1= wird gerade überprüft, 2=fragliche ID, 3=falsche Laborwerte etc.), was die Korrektur erleichtern kann. Bei berechtigten Zweifeln kann der Auswerter dieses Feld wieder auf einen Fehlerstatus setzen, die Datensätze brauchen nicht ständig herumgeschoben werden. Solche Statusfelder können auch für einzelne wichtige oder theoretisch sogar für alle Items verwendet werden. Letzteres würde allerdings den Umfang der Datenbanken verdoppeln! Alternativ könnte man auch verschiedene *Missing*-Codes verwenden (z.B. negative Zahlen); dann ergibt sich allerdings der Nachteil, dass der ursprüngliche (falsche) Wert nicht gespeichert werden kann.

8.6. Datenkorrektur

Eine spätere Korrektur der Daten muss immer möglich sein. Das klingt zunächst banal, ist aber ein Problem, wenn Datensätze mehrfach umgeformt werden. Wenn Sie ihre Fragebögen zunächst in *Access* erfassen, mit Hilfe einer *Abfrage* aus mehreren Tabellen Ihre Matrix für die Auswertung erstellen und diese schließlich via *Excel* nach *SPSS* exportieren, haben Sie schnell etliche Versionen Ihrer Daten in verschiedenen Dateiformaten. Stoßen Sie dann bei der Auswertung auf ein paar unmögliche Werte, die eine Korrektur der Daten erfordern, so wird es problematisch: Wenn Sie den Fehler nur in der Datei verbessern, die Sie gerade bearbeiten, hinterlassen Sie einige unkorrigierte Versionen Ihrer Daten. Und wenn Sie dann noch mehrere Fehler in verschiedenen Dateien verbessern, ist kein File mehr auf dem neuesten Stand. Müssen die Daten in ein anderes Format umgewandelt werden (z.B. von Access nach SPSS), gibt es mehrere Möglichkeiten:

- Sie korrigieren die Werte in der Datenbank und wiederholen die gesamte Umwandlungsprozedur. Das ist das sicherste, aber auch das aufwendigste Verfahren. Diese Prozedur sollte so weit wie möglich automatisiert werden (Batchdatei, Makro, Programmiersprache). Im günstigsten Fall brauchen Sie dann nur die Kette anzustoßen und alles Übrige läuft automatisch ab.
- Sie korrigieren die Fehler sowohl in der Datenbank als auch in Ihrem SPSS-File. Alle eventuell vorhandenen Zwischenfiles (z.B. Excel) werden entweder korrigiert oder gelöscht. Nur so können Sie sicher sein, dass keine Dateien mit fehlerhaften Daten mehr existieren. Die Gefahr liegt darin, einzelne Dateien oder Werte zu vergessen oder durch Tippfehler nicht übereinstimmende Dateien zu erzeugen.
- Dateien, die als Zwischenlösung dienen, sollten so bald wie möglich wieder gelöscht werden. Oder Sie geben ihnen einen Namen, der auf den temporären Charakter hinweist (z.B. Temp001.XLS) bzw. Sie legen sie gleich in einem Ordner ab, der nur temporäre Dateien enthält. Sie dürfen aber nicht die Extension ändern, da *Windows* sonst den Dateityp nicht mehr richtig erkennt.
- Der Idealfall sieht so aus, dass nur ein gültiges Exemplar der Datendatei existiert, etwa in Form einer Datenbank. Alle anderen Programme holen sich die Daten direkt dort ab (Stichwort: dynamischer Datenaustausch, DDE oder ODBC). Dafür müssen aber alle beteiligten Programme diese Verfahren unterstützen.
- Besondere Sorgfalt ist gefordert, wenn die Daten durch mehrere Hände gehen, weil z.B. für die verschiedenen Aufgaben wie Data Management, Auswertung mit Standardsoftware oder die Lösung komplexer

Probleme mit Spezialprogrammen verschiedene Personen eingesetzt werden. Es darf auf keinen Fall passieren, dass jemand seine ‚eigenen' Daten korrigiert, während alle anderen mit dem unveränderten Datensatz weiterarbeiten.

- Viele Statistikprogramme, darunter auch SPSS, berechnen auf Wunsch neue Größen und legen sie in der Datenmatrix ab. Ändern sich Operanden, so hat das oft keine Wirkung auf die berechneten Größen. Gegebenenfalls müssen die Berechnungen also wiederholt werden, damit die Ergebnisse stimmen.
- Die vorgenommenen Datenkorrekturen müssen gut dokumentiert werden. Das ist bei überwachten Studien unbedingt erforderlich, sonst kann die ganze Arbeit abgelehnt werden. Aber auch für Sie selbst ist eine gute Dokumentation wichtig, damit Sie immer wissen, welche Werte Sie in welchen Dateien verändert haben.

9. Einsatz von Datenbanken

9.1. Vor- und Nachteile einer Datenbank

Der Sinn von Datenbanken liegt darin, Informationen zu sammeln und vor allem gezielt wiederzufinden, d.h. Daten zu organisieren (*Data Base Management System* ***DBMS***). Das können Adressen und Telefonnummern, Lagerbestände, Unterlagen für die Abrechnung mit einer Krankenkasse oder auch Daten für eine klinische Studie sein. Im Prinzip handelt es sich nur um die elektronische Form eines Karteikastens. Und wie eine schlecht geführte Kartei die Arbeit nicht unbedingt erleichtern muss, kann auch eine schlecht organisierte Datenbank schnell zu einem Datensumpf werden, aus dem keine brauchbare Information mehr herauszuholen ist. Je komplizierter die Struktur der Daten ist, umso gründlicher muss der Einsatz einer Datenbank geplant werden.
In kleineren Studien ist die Datenerfassung kein großes Problem, vorausgesetzt, der Fragebogen wurde sinnvoll und computergerecht erstellt. Im einfachsten Fall werden die Daten direkt in das Auswertprogramm eingegeben. Bei komplexen Daten kann die Verwendung einer professionellen relationalen Datenbank sinnvoll und notwendig sein. Diese Programme bieten folgende Vorteile:

- Der Fragebogen kann auf dem Bildschirm abgebildet und mit Zusatzinformationen ergänzt werden.
- Bei der Eingabe über eine Datenmaske ist eine individuelle Reaktion auf spezielle Werte möglich, z.B. eine Fehlermeldung, das automatische Auffüllen und Überspringen anderer Felder oder der Wechsel in eine andere Maske.
- Verschiedene Fragebögen (z.B. Anamnese, Laborbogen, Telefoninterviews, Patiententagebücher) werden in verschiedenen Masken zunächst so erfasst, wie sie anfallen und erst später zusammengeführt. Dabei ist es aber unabdingbar, dass jeder Fall eindeutig identifiziert werden kann. Und auch das Problem mit der rechteckigen Datenmatrix ist nur aufgeschoben, aber auf keinen Fall aufgehoben!
- Der Datenschutz ist bei guten Datenbanken höher als bei anderen Programmen, Zugriffsrechte können individuell vergeben werden. Mehrere Personen können gleichzeitig Daten eingeben; das Programm achtet darauf, dass dabei keine Konflikte entstehen (z.B. wenn zwei Anwender versuchen, gleichzeitig den gleichen Datensatz zu verändern). Moderne Datenbanken unterstützen den Zugriff auf die Daten über Netzwerke.

Die Verwendung eines Datenbankprogramms ist sinnvoll, wenn einige der oben genannten Funktionen die Handhabung der Daten deutlich erleichtern. Aber trotz flotter Reklamesprüche der Softwareindustrie muss man sich gründlich in diese Programme einarbeiten. Und auch dann dauert es seine Zeit, bis eine Maske erstellt ist. Individuelle Wünsche müssen auch individuell formuliert, jeder Validitätscheck einzeln definiert und anschließend gründlich getestet werden.
Für die Auswertung sind die Daten aus der Datenbank in das Format des Statistikprogramms umzuwandeln (wichtig: was passiert mit fehlenden Angaben?). Falls mehrere Tabellen existieren, müssen diese entweder mit Hilfe des Datenbankprogramms vorher sinnvoll zu einer einzigen rechteckigen Datei verknüpft werden oder diese Arbeit muss das Statistikprogramm leisten (das meistens mehr Variablen verwalten kann; *Access* erlaubt nur 256 pro Tabelle). Das ist nicht so einfach, wie es sich vielleicht anhört: Zunächst müssen eindeutige Regeln festgelegt werden, nach denen die Verknüpfung erfolgen soll, danach müssen diese in die Sprache des Programms umgesetzt werden, wobei manchmal die Struktur der Daten komplett verändert werden muss. Bei komplexen multizentrischen Studien ist nicht selten ein Spezialist nötig, der sich in Vollzeit ausschließlich um die Verwaltung der Dateien kümmert.

9.2. Allgemeines

Alle besseren Datenbanken bieten die Möglichkeit, individuelle Tabellen zu definieren, Eingabemasken zu erstellen und einfache Abfragen und Reports zu generieren. Für komplexere Fragen bieten viele Programme eine eigene Programmiersprache an. Das heißt, typische Datenbankfunktionen sind als Standard verfügbar, für Spezialfälle müssen dann individuelle Programme entwickelt werden.

Die folgenden Punkte sollte eine gute Datenbank bieten:

- Gute Dokumentation (gedruckt) *und* gute Online-Hilfe.
- Definition von Tabellen mit frei zu definierenden Feldern in verschiedenen Formaten (incl. Datum).
- Umgang mit fehlenden Angaben.
- Einfache visuelle Generation von Eingabemasken.
- Prüfen von doppelt eingegebenen Datensätzen auf Übereinstimmung *).
- Zugriff durch mehrere Benutzer gleichzeitig *).
- Zugriffsrechte auf verschiedenen Ebenen *).
- Die typischen *Windows*-Funktionen wie Mausbedienung, Auswahlfelder, Dialogboxen oder Hilfe sollten ohne großen Aufwand zur Verfügung stehen.
- Die Programmierschnittstelle sollte direkt auf den vorhandenen Standards aufsetzen, d.h. es müssen nur für die Sonderwünsche spezifische Programme erstellt werden.

- Bei überwachten Studien wird manchmal verlangt, dass die Datenbank automatisch alle Eingaben und Änderungen überwacht und protokolliert *).

*) Einige Elemente bieten nur teure Spezialprogramme, nicht die Datenbanken, die zu Office-Paketen gehören (z.B. *MS Access*).

9.3. Die Elemente einer relationalen Datenbank

Am häufigsten werden Sie auf sogenannte *Relationale Datenbanken* treffen. Die Grundelemente bilden auch hier rechteckige Tabellen, von denen Sie aber mehrere anlegen und miteinander verknüpfen können.

9.3.1. Tabellen

Die Grundlage bilden rechteckige Tabellen, in denen die Spalten die Merkmale (***Felder***) und die Zeilen die Einträge (***Datensätze***) darstellen. Der Unterschied zu Programmen wie SPSS liegt darin, dass mehrere solcher Tabellen verknüpft werden können.

Beispiel: Die Basisdaten für jeden Patienten (Alter, Geschlecht, Anamnese usw.) bilden eine Tabelle, in der jedes Individuum genau einmal vorkommt. In einer zweiten Tabelle „Laborbogen“ werden die Ergebnisse der Untersuchungen aufgenommen. Die Tabelle selbst ist ebenfalls rechteckig (es werden immer die gleichen Parameter gemessen), jedes Individuum kann aber mehrfach in der Tabelle vorkommen, die Einheit ist also nicht der Patient, sondern die Untersuchung. Dass dies bei statistischen Auswertungen zu einem verzerrten Bild führen kann, wurde schon erwähnt. Zu einem spezifischen Datensatz ***(Record)*** in den Basisdaten können also kein, ein oder mehrere Sätze in der Tabelle „Laborbogen“ gehören (1: n).

Zuerst müssen Sie die Tabellen anlegen und die Felder definieren (Name und Datentyp; je nach Datenbank auch Nachkommastellen, erlaubte Werte, Pflichtfelder, ein- oder mehrdeutiger Index). Anschließend können Sie Ihre Werte direkt in die Tabelle eingeben, dabei prüft das Programm automatisch den Datentyp und - soweit definiert - ob ein erlaubter Wert und ein gültiger Index vorliegt. Weitergehende Checks erfordern meistens eine spezielle Eingabemaske.

Auch in einer Datenbank muss jede einzelne Zeile eindeutig identifizierbar sein. Besonders wichtig ist die gemeinsame Patientenidentifikation, da das Programm ausschließlich diesen Schlüssel benutzt, um die Labordaten den Patienten in der Anamnesetabelle zuzuordnen. Während die Patientennummer in der Basistabelle ausreicht, um eine Zeile eindeutig zu bezeichnen (jeder Patient kommt nur einmal vor), muss in der Labortabelle der Schlüssel erweitert werden, z.B. indem die Untersuchungen pro Patient durchnumeriert werden. Statistikprogramme können allerdings mit solchen Datenstrukturen meist nicht viel anfangen. Für die Auswertung muss aus diesen Tabellen wieder eine rechteckige Matrix erzeugt werden, in der die Untersuchungen für einen Probanden nebeneinander stehen. Einfacher und weniger fehleranfällig ist es in jedem Fall, die Daten von vornherein ‘rechteckig’ zu planen, also schon vor der Studie festzulegen, wann welche Werte bestimmt werden sollen.

9.3.2. Eingabemasken

Eingabemasken sollen die Dateneingabe erleichtern, indem nicht nur Platz für die Eingabe der jeweiligen Werte zur Verfügung steht, sondern auch weitere Informationen auf dem Bildschirm erscheinen. Bei Benutzung von Eingabemasken ist es oft wünschenswert, dass die Formulare auf Papier und auf dem Bildschirm gleich aussehen.

Abgesehen von der schöneren Optik und der besseren Übersicht sollte eine Datenbankmaske einige zusätzliche Funktionen zur Verfügung stellen:

- Alle typischen *Windows*-Elemente wie Ankreuzfelder, Optionsfelder oder Auswahllisten sollten einfach zu realisieren sein.
- Das Programm sollte in der Lage sein, auf bestimmte Ereignisse wie z.B. das Ändern einer Eintragung, das Anklicken einer Schaltfläche oder den Aufruf eines anderen Datensatzes individuell zu regieren, d.h. es muss möglich sein, bei Bedarf jedem solchen Ereignis ein eigenes Programm zuzuordnen.

Zu den wichtigsten Aufgaben einer Eingabemaske gehört, Fehler sofort zu erkennen und zurückzuweisen. Einige dieser Fehler werden auch erkannt, wenn die Eingabe direkt in der Tabelle erfolgt.

- ***Pflichteingabe:*** Meldet einen Fehler, wenn Felder nicht ausgefüllt werden. Felder, die der Indizierung dienen, sind immer Pflichtfelder, da eine Datenbank die Datensätze sonst nicht einordnen kann. Werden im Extremfall nur Pflichtfelder angelegt, so garantiert das die 100%ige Vollständigkeit der Daten. Das bedeutet aber auch, dass die Aufnahme von fehlenden Werten grundsätzlich verweigert wird, d. h. das Fehlen eines Items bedingt den Verlust des ganzen Datensatzes.
- ***Eindeutiger Index:*** Das Programm speichert keinen Datensatz, wenn der eingegebene Wert für ein als eindeutig definiertes Schlüsselfeld bereits in der Datenbank vorkommt.
- ***Datentypprüfung*****:** Wenn einem Feld ein spezieller Datentyp zugewiesen wird (ganze Zahl, Dezimalzahl, Datum, Zeitangabe, Text), so überprüft das Programm die korrekte Eingabe und zeigt eventuelle Fehler sofort an.

- ***Bereichsprüfung:*** Für jedes Item werden bestimmte Minima und Maxima oder eine Liste von erlaubten Werten vorgegeben, die Aufnahme von Werten außerhalb dieser Vorgabe verweigert. Dies verringert Kommafehler bzw. vergessene Ziffern (100 statt 1000) und verhindert möglicherweise, dass Daten in das falsche Feld eingegeben werden. Bei codierten Antworten ist diese Methode besonders sinnvoll, wenn z.B. bei Schulnoten alle Werte außer 1, 2, 3, 4, 5 oder 6 zurückgewiesen werden. Sie müssen allerdings bei der Definition aufpassen, ob halbe Noten (drei bis vier = 3,5) möglich sind. Dann müssten Sie 1 als Minimum und 6 als Maximum setzen und Sie dürfen kein Integer-Feld verwenden. Wenn die Zahl der möglichen Antworten beschränkt ist, kann die Eingabe auch gleich aus einer vorgegebenen Liste ausgewählt werden. Beachten Sie aber, dass nicht jeder davon begeistert ist, jedes Mal per Maus oder Cursortasten eine Antwort auszuwählen, statt gleich mit einem einzigen Tastendruck den Code einzugeben. Wenn schon, dann benutzen Sie Kombinationsfelder, die beide Möglichkeiten erlauben.
  Bei gemessenen Zahlen wie Laborwerten müssen die Grenzen so weit gefasst sein, dass auch extreme Werte eingegeben werden können, falls sie vorkommen. Die Maschine Computer würde die Eingabe korrekt gemessener Werte sonst rigoros verweigern. Eine gute Möglichkeit wäre es, absolute und plausible Grenzen vorzugeben. Liegt ein Ausreißer außerhalb der Plausibilitätsgrenze, erscheint eine Warnung, die aber per Mausklick die Aufnahme des Wertes ermöglicht. Leider gibt es diese Möglichkeit in Datenbanken sehr selten. Um sie zu realisieren, muss die Programmiersprache der Datenbank eingesetzt werden, was wieder mehr Aufwand bedeutet.
- ***Plausibilitätskontrolle zwischen Items*:** Die Angaben „Geschlecht = männlich“ und „Zahl der Geburten = 3“ sind beide für sich gesehen möglich, als Kombination aber eher unwahrscheinlich. Solche feldübergreifenden Kontrollen müssen in der Regel über das mit der Datenbank gelieferte System (z.B. Visual BASIC for Applications, VBA) programmiert werden. In jedem Fall müssen für jedes Feld alle möglichen Widersprüche überlegt und definiert sowie die Reaktion im Fall des Falles festgelegt werden. Ein Aufwand, der sich bei kleinen Datenmengen kaum lohnt, bei größeren Projekten aber Zeit sparen kann. Unter Umständen können auch die Grenzen für ein Feld vom Inhalt eines anderen Feldes abhängig gemacht werden (z.B. geschlechtsspezifische Laborwerte wie Hämoglobin, oder die Körpertemperatur bei verschiedenen Krankheiten). Eine Variante dieses Verfahrens ist es, bei Eingabe bestimmter Werte andere Items automatisch auszufüllen oder zu überspringen (z.B. die Fragen nach einzelnen Beschwerden, wenn der Patient sich für beschwerdefrei erklärt hat).
- ***Plausibilitätskontrolle zwischen Tabellen*:** Im Prinzip das gleiche wie oben, aber hier werden Items aus verschiedenen Tabellen (d.h. verschiedenen Fragebögen) abgeglichen. Beispielsweise könnte das Programm in den Stammdaten nachschlagen, ob die im Laborbogen eingegebene ID-Nummer schon vorhanden ist und ob die Initialen des Patienten stimmen. Dies erfordert aber, dass das Datenbankprogramm mehrere Tabellen gleichzeitig offen halten kann, und auch der Aufwand beim Programmieren wird größer. Außerdem entstehen Wartezeiten bei der Eingabe.

9.3.3. Abfragen

Unter Abfragen versteht man die Möglichkeit, anhand von vorgegebenen Bedingungen aus einer oder mehreren Tabellen Daten herauszusuchen. Mit diesem Werkzeug können Sie also aus mehreren Tabellen einer relationalen Datenbank die für die statistische Analyse benötigte rechteckige Datenmatrix erstellen. Moderne Datenbanken helfen ungeübten Benutzern über *„Assistenten“* bei der Zusammenstellung einer formal korrekten Anweisung. Diese helfen Ihnen, die Bedingungen in der Sprache der Datenbank zu formulieren. Die Bedingungen müssen Sie aber selbst aufstellen, und zwar in der für Computer benötigten Präzision. Manche Datenbanken schränken die Zahl der erlaubten Felder (Variablen) pro Tabelle bzw. Abfrage ein (*Access 2003* kann z.B. immer noch nicht mehr als 256 Felder verwalten!).

9.3.4. Reports

Reports sind Berichte, die in der Datenbank enthaltene Informationen zusammenfassen. Für einfache Standardaufgaben wie Auszählungen oder Berechnung von Mittelwerten existieren in der Regel fertige Prozeduren, ansonsten müssen individuelle Programme geschrieben werden. Für statistische Analysen sind diese Reports in der Regel nicht geeignet.

9.3.5. ODBC und SQL

ODBC steht für ***Open Data Base Connectivity*** und ist eine Möglichkeit, um Daten zwischen verschiedenen Programmen auszutauschen. Die diversen Datenbanken versuchen sich in Funktionalität und Komfort gegenseitig zu übertreffen und bieten - wie auch bei anderen Programmen üblich - jeweils eine eigene Benutzersprache und -oberfläche an. Es gibt aber die Abfragesprache **SQL** (***Structured Query Language***), die jedes bessere Datenbankprogramm beherrschen sollte. Mit SQL kann man also (im Prinzip) aus jeder Datenbank Daten auslesen, aber natürlich nicht alle Möglichkeiten der Datenbank ausnutzen. *ODBC* setzt auf *SQL* auf.
Damit ODBC funktioniert, müssen sich die entsprechenden Programme als *Server* (Datengeber) und/oder als *Client* (Datenempfänger) im System anmelden. Das geschieht normalerweise automatisch bei der Installa-

tion. SPSS kann u.a. Daten von MS Access oder auch von MS Excel bekommen. Wenn Sie Daten von Excel via ODBC einlesen, können Sie sogar Formeln stehen lassen, SPSS bekommt automatisch die Werte. Wenn die entsprechenden Treiber installiert wurden (***MySQL ODBC***, natürlich nicht in Windows enthalten), können z.B. auch Daten gelesen werden, die auf einem entfernten Linux-Server abgelegt sind.
Ein Nachteil ist, dass die Feldnamen in SQL andere Beschränkungen haben als in SPSS. Sowohl Access wie auch SPSS erlauben Namen wie AlkPh.1. In SQL ist ein Punkt im Namen aber nicht erlaubt, also auch nicht in ODBC. In der Regel gehen deshalb keine Variablen verloren, der Name wird aber automatisch geändert (z.B. in AlkPh1). Folglich findet SPSS anschließend die Variable AlkPh.1 nicht.

10. Statistiksoftware

Es gibt auf dem Markt etliche Programme zur Auswertung statistischer Daten. Neben mehreren großen Paketen existieren auch Programme für Spezialaufgaben, die in den großen Paketen nicht enthalten sind. Immer wieder kaufen die großen Pakete (die in der Regel auch eigene Firmen sind) kleinere Firmen auf und verleiben sich deren Software bzw. Algorithmen ein.
Die Programme werden entweder über eine spezielle Kommandosprache gesteuert oder über eine graphische Oberfläche, wie man sie von Windows kennt *(GUI = Graphical User Interface)*. Die Klickoberfläche ist natürlich einfacher zu handhaben und schneller zu lernen, die Kommandosprache erlaubt es dagegen, Abläufe zu programmieren und abzuspeichern und so zu automatisieren. Manche Pakete bieten beide Möglichkeiten an.

10.1. SPSS

Ursprünglich „Statistical Package for Social Scientists“, also für Soziologen geschrieben. Es beinhaltet alle gängigen statistischen Verfahren und erzeugt als Ausgabe formatierte Tabellen sowie Grafiken. Die Befehle können zusammengeklickt werden oder über die sogenannte „Syntax“ programmiert ablaufen. Gut gelungen ist die Verbindung der beiden Oberflächen, auf Wunsch erzeugt SPSS für zusammengeklickte Befehle die korrekten Syntaxbefehle. Die Bedienung ist relativ einfach, dadurch ist SPSS auch für Einsteiger gut geeignet.
Früher war SPSS inc. eine eigene Firma, im Herbst 2009 wurde das Programm von IBM übernommen. Mit der Übernahme wurde auch die vorübergehende Umbenennung in *PASW* rückgängig gemacht. Der offizielle Name ist jetzt **IBM SPSS Statistcs**. Die Charité verfügt über eine Netzwerklizenz; d.h. berechtigte Benutzer kommen günstig an das Programm.

10.2. SAS

Der größte Konkurrent von SPSS und noch mächtiger im Umfang. Wird besonders von großen Konzernen eingesetzt wie Banken oder Versandhäusern. Speziell bei der Transformation von Daten bietet es mehr Möglichkeiten als SPSS. Die Komplexizität des Programms erschwert aber auch die Benutzung. Es gibt zwar so etwas wie eine GUI, aber sinnvollerweise benutzt man nur die Kommandosprache. Die Einarbeitung dauert länger. Auch hier gibt es eine Campuslizenz.

10.3. S-Plus / R

Ursprünglich sollte eine „statistische Programmiersprache“ auf der Basis der Sprache „C“ entwickelt werden. Diese bekam den Namen „S“. Später wurde das Programm mit einer graphischen Oberfläche versehen und als „S-Plus 2000“ kommerziell vertrieben. Alternativ wird es als weiterhin kostenlose Version unter dem Namen „R“ angeboten und weiterentwickelt. „R“ darf - ähnlich wie Linux - unter Beachtung der Bedingungen kostenlos heruntergeladen, verwendet oder verbessert werden **(http://www.r-project.org/)**. Ohne Grundkenntnisse in Mathematik (Vektoren und Matrizen) und Informatik (Programmieren) fällt die Benutzung schwer. Dafür ist es möglich, das Paket durch selbst geschriebene Auswertungsprogramme zu erweitern. SPSS bietet die Möglichkeit, R-Programme einzubinden.

10.4. Stata

Stata ist ein weiteres umfassendes Statistikpaket und in mehreren Versionen für Windows, Mac und Linux erhältlich. Es bietet Grafik, eine GUI und eine eigene Programmiersprache. Großabnehmern und Universitäten werden zwar Rabatte eingeräumt, eine Campus-Lizenz gibt es aber nicht.

11. Datensicherung

Dass man von allen wichtigen Daten Backups anfertigen sollte, weiß eigentlich jeder. Die in einem Computer gespeicherten Daten können verloren gehen, manchmal sogar der ganze PC. Möglichkeiten dazu gibt es mehr als genug. Angefangen von Feuer und Diebstahl, defekten Festplatten, beim Transport beschädigten oder verlorenen Datenträgern über Systemabstürze, die Dateien beschädigen und bösartige Software (Viren, Würmer & Co) bis hin zu versehentlich selbst gelöschten Dateien ist alles möglich. Zur Sicherheit sollten Sie immer Kopien Ihrer Daten auf mehreren unabhängigen Datenträgern anfertigen. In der Praxis werden sie allerdings schnell vergessen. Da arbeitet man am Freitag intensiv an einer Sache und merkt plötzlich, dass man schon zu spät zu einem Termin kommt. Schnell alles speichern und den Rechner herunterfahren. Kaum zu begreifen, wenn der PC dann am Montagmorgen erklärt, er habe keine Festplatte gefunden! Aus nicht erklärbaren Gründen war die 4 Monate alte Platte defekt, kein Trick (in anderen PC einbauen, austauschen der Elektronik) konnte sie reanimieren und die Daten retten. Da ist die Erklärung, dass die Platte selbstverständlich auf Garantie ersetzt wird, nur ein kleiner Trost. Verlorene Daten kann niemand wieder herbeizaubern! Es gibt zwar Firmen, die eine Wiederherstellung der Dateien von defekten Festplatten anbieten, Wunder vollbringen können diese aber auch nicht und die Dienste sind teuer. Wichtige Aspekte beim Backup sind Sicherheit, Komfort, Umfang der Daten, Schnelligkeit und natürlich die Kosten.
Die Frage ist, was mit der Sicherung erreicht werden soll. Bevor ich wesentliche Änderungen an einer Datei vornehme, mache ich meistens eine Kopie und ergänze den Dateinamen um das Datum (z.B. nenne ich die ‚eingefrorene' Version der Datenbank „Hydro" vom 15. 6. 2002 Hydro-020615.mdb, Hydro.mdb ist immer die aktuelle Version). Dann kann ich notfalls einen früheren Zustand reaktivieren. Wenn allerdings die Festplatte ihren Geist aufgibt, sind auch diese Kopien verloren. Man könnte eine zweite Festplatte (z.B. aus einem alten PC) einbauen und darauf alle wichtigen Daten doppeln. Das hilft gegen einen Plattencrash, nicht aber bei Brand oder Diebstahl. Kopien über ein Netzwerk auf einer entfernten Festplatte sind besser. Aber alle Platten, auf die ich immer direkten Zugriff habe, stehen leider auch Viren und anderen digitalen Schädlingen offen. Also muss ich die wichtigen Daten auf externe Datenträger (Platte, Stick) schreiben und diese sicher verwahren, möglichst weit vom PC entfernt (Brand). Da so eine Sicherung aber nicht gegen versehentlich gelöschte oder defekte Dateien sowie Viren & Co schützt (alle Fehler werden mitgesichert), sollte man mehrere Sicherungen aufheben, etwas veraltete Daten sind besser als gar keine. Wie oft man sichert und welcher Aufwand dabei betrieben wird, hängt von der Wichtigkeit und Brisanz der Daten und nicht zuletzt vom finanziellen Etat ab (professionelle Backupsysteme kosten weit mehr als ein gut ausgestatteter PC). Es ist sicher ärgerlich, wenn sich die Word-Datei mit Ihrem Vortrag in Luft auflöst, Sie können aber den ausgedruckten Text in 1-2 Stunden wieder eintippen. Ob es sich lohnt, stattdessen täglich mehr als eine halbe Stunde in Datensicherungen zu investieren, ist fraglich. Anders sieht es aus, wenn von den Daten die Existenz einer Firma oder einer kostenaufwendigen klinischen Studie abhängt. Hier kann sich auch die Anschaffung teuer Hard- und Software lohnen. Über Strategien beim Sichern kann man ganze Bücher füllen. Eine universelle Methode, die für jeden die optimale Lösung darstellt, gibt es nicht. Bei Ihren Überlegungen sollten Sie folgende Aspekte berücksichtigen:

11.1. Typ der Sicherung

Dateibasierte Sicherung: Das System liest das Inhaltsverzeichnis der Quelle, kopiert alle dort aufgeführten Dateien nacheinander auf das Ziel und erstellt dort ein neues Inhaltsverzeichnis. *Vorteil:* Auch wenn durch häufiges Umkopieren die Teile (Sektoren) der einzelnen Dateien über die ganze Platte verstreut sind, liegen sie bei der Kopie direkt hintereinander. Außerdem werden nur vorhandene Dateien kopiert; wenn nur wenig auf der Platte drauf ist, geht das Kopieren entsprechend schnell. *Nachteil:* Versteckte Strukturen werden nicht erkannt und auch nicht kopiert. Bei normalen Dateien, die Sie z.B. mit *Word* oder *SPSS* erzeugt haben, spielt das keine Rolle. Handelt es sich aber um eine bootfähige CD oder Festplatte (z.B. das Betriebssystem oder von einem Virenscanner erstellte Rettungsdisketten), wird die Kopie nicht funktionieren.
Sektorbasierte Sicherung: Hier wird die Quelle Sektor für Sektor ausgelesen und genau so auf das Ziel geschrieben, es entsteht also ein genaues Abbild (ein ***Clone***) des Originals. Vorausgesetzt, Laufwerk und Computer können die Daten lesen, was man manchmal mit diversen absichtlich eingebauten Fehlern verhindern will (beim normalen Backup von Daten auf der Festplatte spielen solche Kopierschutzmechanismen keine Rolle). Viele Programme, die auf das Klonen von Datenträgern spezialisiert sind, legen auf Wunsch interne Zwischendateien an, die als ***Image*** (eine Art Negativ) bezeichnet werden. Es handelt sich um eine ganz normale Datei, die aber meistens von keinem anderen Programm lesbar ist. Das Clone-Programm kann aber mit Hilfe dieser *Image-Datei* wieder exakte Abbilder des Originaldatenträgers erstellen.
Sie können z.B. auf Ihren PC Windows neu installieren, alle vorhandenen Updates einspielen, die neuesten Treiber für Ihre Hardware herunterladen und noch ein paar Hilfsprogramme installieren, ohne die Sie nicht mehr arbeiten wollen. Das Ganze kann einige Stunden dauern. Erstellen Sie dann ein Image von der Systemplatte (C:\) und heben Sie es gut auf. Im Fall des Falles können Sie dann diesen Zustand wieder

herstellen, dabei geht aber der vorhandene Inhalt dieser Festplatte verloren. (Ein solcher Datenträger liegt gekauften PCs oft als „Recovery-CD“ bei). *Nachteil*: Ziel und Quelle müssen physikalisch gleich und gleich groß sein (Image-Dateien können allerdings auf einem beliebigen Datenträger abgelegt werden). Es wird immer die ganze Platte kopiert, auch wenn nur wenige Sektoren sinnvolle Daten enthalten, außerdem gehen auf dem Zieldatenträger natürlich alle eventuell vorhandenen Daten verloren.

11.2. Software

Sie können für Ihre Sicherungen einfach den Explorer verwenden, indem Sie die entsprechenden Dateien und Verzeichnisse auf einen anderen Datenträger kopieren. Sie müssen aber selbst die Übersicht behalten, was Sie wann wo und wie gesichert haben. Um Platz zu sparen, können Sie die Daten in Archive verpacken (ZIP, 7zip, RAR). Dabei werden alle ausgewählten Verzeichnisse und Dateien in eine einzige Datei zusammengefasst und zusätzlich - soweit möglich - komprimiert. Einige verteilen das Archiv bei Bedarf automatisch auf mehrere Datenträger (*Multi-Volume*) und ersparen Ihnen so einiges an Rechnerei. Um aus so einem Archiv wieder die Originaldaten zu gewinnen, brauchen Sie wieder das Archivierungsprogramm. Welches das ist, erkennen Sie an der Dateiendung. Wichtig ist der verwendete Algorithmus, nicht die Oberfläche. So können Sie z.B. mit WinZIP alle ZIP-Archive auspacken, auch wenn sie z.B. unter Linux angelegt wurden. Erscheint ein Archiv als .EXE-Datei, so ist die Routine zum Auspacken enthalten, Doppelklick genügt (Vorsicht bei unklarer Herkunft, es könnte auch schädliche Software sein!).
Laufwerken, die sich zur Datensicherung eignen, ist oft ein Backup-Programm beigelegt. Dieses ist meistens auf diesen Datenträger spezialisiert, kann also mit fremder Hardware wenig anfangen. Auf den verschiedenen *Windows*-CDs sind verschiedene (und untereinander meist nicht kompatible) Backup-Programme. Um sie zu installieren, muss ein benutzergesteuertes Setup durchgeführt werden.
Die Alternative sind professionelle Backup-Programme. Diese legen die Daten fast immer in einem eigenen Format ab. Dabei versuchen sie durch Komprimierung Platz zu sparen, und wenn der Platz nicht reicht, werden automatisch weitere Datenträger angefordert. In einer speziellen Datenbank wird festgehalten, wann welche Dateien auf welchem Medium gespeichert wurden. Zur Wiederherstellen der Daten ist das passende Restore-Programm erforderlich. Die globale Wiederherstellung (nach einem Totalverlust alle auf dem Datenträger gespeicherten Dateien wieder auf die Festplatte zurückspielen) ist meistens problemlos. Gezielt einige ausgesuchte Dateien zurückzuholen (z.B. weil sie versehentlich gelöscht wurden), ist aufwendiger. Ein Nachteil ist, dass Sie auf das Programm angewiesen sind. Wenn die Software auf dem neuen Sysem nicht mehr funktioniert, müssen Sie das Upgrade kaufen. Sollte es das Programm nicht mehr geben, sind Ihre Sicherungen wertlos. Ungeeignet sind Programme, die durch einen Kopierschutz an eine bestimmte Hardware gebunden sind.

11.3. Hardware

Welche Hardware für die Sicherung zu verwenden ist, hängt von der anfallenden Datenmenge und vom Geldbeutel ab. Es gibt Systeme, in denen Bandkassetten mit hoher Kapazität wie in einer Musicbox bei Bedarf automatisch zum Laufwerk gebracht und geschrieben oder gelesen werden können. Über das lokale Netzwerk werden angeschlossene PCs in der Nacht automatisch eingeschaltet und die Daten gesichert. Das Ganze kostet dann aber oft mehr als alle angeschlossenen Computer zusammen.

11.3.1. Speichermedien

11.3.1.1. Magnetbänder

Magnetbänder waren zur Zeit der Großrechner (Mainframes) das meistbenutzte Speichermedium für Archiv und Transport. Die Technik war vom Tonbandgerät bereits bekannt und Bänder verhältnismäßig billig. Für den Heim- und Anwenderbereich gab es Kassettengeräte (***Streamer***), wobei handelsübliche Kompaktkassetten und DAT (Digital Audio Tape) verwendet wurden, aber auch diverse speziell für die EDV entwickelte Systeme (Travan, Ultrium). In jedem Fall muss ein passendes Laufwerk angeschafft werden, das sonst zu nichts nütze ist, außerdem ist für heutige Maßstäbe der Preis hoch, die Kapazität niedrig und ein „Bandsalat“ kann die gespeicherten Daten schnell vernichten. Passte früher der Inhalt mehrerer Festplatten auf ein Band, braucht man heute mehrere Bänder zum Sichern einer Festplatte. Wie bei Audio und Video haben Bandgeräte im SoHo-Bereich (Small Office / Home Office) keine Zukunft mehr.

11.3.1.2. Disketten

Disketten funktionierten im Prinzip wie ein Tonbandgerät, nur dass statt des Bandes eine Scheibe verwendet wurde. Das hat den Vorteil, dass jede Stelle ohne langes Spulen zu erreichen ist. Der Kopf wird beim Lesen und Schreiben fest an die Magnetscheibe gedrückt, die nachgeben und sich an den Kopf anschmiegen muss (*Flexible Disk* oder auch *Floppy Disk*). Das macht höhere Geschwindigkeiten natürlich unmöglich. Die ersten Disketten hatten eine Größe von 8 Zoll, die ersten PCs liefen mit Minidisketten (5¼") und später mit den robusteren Mikrodisketten (3½"), die zuletzt 1,44 MB speichern konnten. Es gab Laufwerke von etlichen Herstellern, aber alle Gehäuse waren gleich groß und hatten die Schraubgewinde an den gleichen Stellen, damit sie in jeden PC eingebaut werden konnten. Diese Formate sind heute noch als Standard in

jedem PC-Gehäuse zu finden. Wenn in einem Prospekt steht, dass ein Gehäuse über vier 5¼"“-Schächte verfügt, können Sie bis zu 4 genormte Laufwerke einbauen, die Disketten in diesem Format lesen können. Heute findet man hier eher optische Laufwerke oder auch ein Frontpanel für die Audiokarte. Es gab diverse Ansätze, die Kapazität der Diskette auf 2,88 MB zu verdoppeln oder neue Diskettensysteme mit mehr Kapazität zu entwickeln (***IOMEGA ZIP*** und ***Superdisk***). Besonders die *ZIP*-Laufwerke , deren Disketten 100, 250 oder 750 MB aufnehmen konnten waren durchaus beliebt, sind aber seit der Einführung der USB-Sticks bedeutungslos geworden und werden nicht mehr hergestellt.

11.3.1.3. Festplatten

Festplatten sind im Prinzip die Weiterentwicklung der Diskette. Der wichtigste Unterschied ist der, dass der Kopf auf einem Luftpolster schwebt und die Platte nie berührt (wenn doch, zerstört das den Kopf, die Plattenoberfläche und somit natürlich auch die Daten). Da es keine Reibung gibt, kann die Platte schneller drehen. Die Platte selbst muss absolut starr und plan sein, damit der Abstand zum Kopf immer gleich bleibt, daher der Name ***Hard Disk***. Schon ein Staubkorn oder ein Rauchpartikel ist um ein vielfaches größer als der Abstand zwischen Kopf und Platte und kann somit einen *Headcrash* auslösen. Deshalb werden Festplatten immer in ein hermetisch abgeschlossenes Gehäuse eingebaut. Die ersten Festplatten mit einer Kapazität von 20MB haben mehr gekostet als ein PC mit 2 Diskettenlaufwerken. Derzeit bezahlen Sie für eine externe Festplatte mit 2 TB (ca. 2 000 000 MB) um die 100 Euro. Gleichzeitig wurden die Laufwerke schneller und zuverlässiger. Für die Sicherung größerer Datenmengen verwendete man früher externe Platten im 3½"-Format. Sie werden in einem stabilen (schweren) Gehäuse und mit Netzteil für die Stromversorgung verkauft. Für den mobilen Einsatz gibt es externe Notebook-Platten im Format 2½-Zoll. Sie wiegen weniger, passen in jede Tasche und brauchen meist keine separate Stromversorgung. Im Preis-Leistungs-Verhältnis sind Festplatten heute unschlagbar, daher sind sie auch als Backup-Medium gut geeignet.

11.3.1.4. Wechselplatten

Das Laufwerk wird in den Computer eingebaut (oder extern angeschlossen), die Information auf auswechselbaren Medien gespeichert. Außer bei Disketten und optischen Laufwerken verwendet man heute kaum noch Wechselplatten, da sie teurer und störanfälliger sind als externe Festplatten.

11.3.1.5. Flash-Speicher

Flash-Speicher ist ein spezieller Speicherbaustein. Während der im PC verbaute Arbeitsspeicher (RAM) ohne Strom sofort alle Information verliert, bleibt der Inhalt des Flash-Speichers erhalten. Dafür benötigt er eine relativ hohe Spannung um beschieben zu werden. Ursprünglich setzte man diesen Speichertyp auf dem PC-Mainboard und in elektronischen Geräten wie Druckern, Recordern oder auch Waschmaschinen ein, um das *BIOS* bzw. die *Firmware* zu speichern. Heute nutzt man diesen Speichertyp vor allem in diversen Speicherkarten und USB-Sticks. Ihr Vorteil ist, dass die Speicher klein und handlich sind, keine Mechanik beinhalten und nur wenig Strom brauchen. Sie werden in Musikplayern und digitalen Kameras eingesetzt. Im Vergleich zu Festplatten sind sie aber teurer und fehleranfälliger, je nach Bauart sind 10.000 bis 2 Millionen Schreibzyklen möglich, danach ist der Chip verbraucht. Defekte Teile werden markiert und nicht mehr verwendet; wodurch die angezeigte Kapazität immer kleiner wird.
Bei den Speicherkarten ist das Problem, dass verschiedene Hersteller unterschiedlich geformte Karten hergestellt haben (CompactFlash CF, Memory Stick MS, Multimedia Card MMC), wirklich durchgesetzt hat sich die Secure Digital Memory Card SD. Diese gibt es in drei verschiedenen Vatianten (SD bis 2 GB, SDHC bis 32 GB und SDXC bis 2048 GB), die abwärtskompatibel sind. In ein Gerät, das nur einfache SD-Karten lesen kann, dürfen keine HC- oder XC-Karten eigelegt werden, das kann sogar zum Datenverlust führen. Während eine SD-Karte etwa die Größe einer Briefmarke hat, gibt es für Kleinstgeräte (Handy, MP3-Player) auch SD-Karten im Mini- und im Mikroformat, die mit Hilfe eines Adapters auch wie eine „normale“ SD-Karte verwendet werden können.
In kleinen Notebooks und Tablets wird Flash-Speicher als „Festplatte“ verwendet, erkennbar an der Bezeichnung ***Solid State Drive*** und der verhältnismäßig geringen Kapazität (oft 256 oder 512 GByte).

11.3.1.6. Optische Laufwerke

Optische Laufwerke richten einen gebündelten Lichtstrahl (Laser) auf eine rotierende Scheibe. Ist die Oberfläche unversehrt, reflektiert sie das Licht auf einen Sensor und das System erkennt eine „1“, sonst wird eine „0“ gelesen. Bei industriell hergestellten Medien (Audio CD, CD-ROM, DVD-Video) werden Vertiefungen (*pits*) eingepresst, die den Laser ablenken. Bei gebrannten Medien verändert ein verstärkter Laser die Oberfläche, damit sie an diesen Stellen nicht mehr reflektiert. Auch normales Licht beschädigt auf die Dauer die Schicht so einer Scheibe, weshalb sie möglichst dunkel aufbewahrt werden sollten. Als Langzeitarchiv (z.B. für Krankenakten, die 30 Jahre aufgehoben werden müssen) sind diese Medien nicht zugelassen. Aber auch die eigene Foto- oder Videosammlung sollte von Zeit zu Zeit überprüft und ggf. umkopiert werden.
Begonnen hat alles mit der Audio CD (bis 700 MB für 74 Minuten), es folgte die DVD (bis zu 2*4,7 GB) und schließlich die Blu-ray Disk (BD, als geschützter Name nicht Blue Ray), die mit einem violetten Laser mit kürzerer Wellenlänge arbeitet und bis zu 2*25 GB speichern kann. Von jedem Typ gibt es einige

Varianten. Die Endung -ROM (Read only memory) wird für gepresste Scheiben verwendet. -R steht für recordable, sie können nur einmal bespielt werden. Bei -RWs ist die Veränderung der Oberfläche reversibel, sie können mehrere Male gelöscht und neu bespielt werden. Schließlich legte man zwei Schichten übereinander, der Laser kann je nach eingestelltem Focus die obere (durchlässige) oder die untere Schicht abtasten (*Double Layer DL).*
Als Archiv für abgeschlossene Projekte sind sie preiswert und geeignet. Es macht aber nicht viel Sinn, jeden Tag mehrere CD-Rs zu brennen und einen Tag später auf den Müll zu werfen. In jedem Fall sollte jede frisch gebrannte CD oder DVD mit dem Original verglichen werden (verify); nur so ist sichergestellt, dass sie keine Fehler enthält (in vielen Brennprogrammen als Option enthalten).
Eine Sonderform ist die DVD-RAM, die wie eine Festplatte organisiert ist, alle Aufzeichnungen werden sofort überprüft. Das macht die DVD-RAM etwas langsamer, aber auch zuverlässiger. Da sie fast beliebig oft überschrieben werden kann, ist sie eine gute Lösung für Sicherungen. Leider sind die Laufwerke selten und die Medien nicht überall erhältlich.

11.3.2. Anschlussmöglichkeiten

11.3.2.1. Interner Einbau

Früher waren vor allem Festplatten von außen unsichtbar fest im PC eingebaut. Wenigstens eine Festplatte findet sich in jedem PC, auf der das Betriebssystem und die benötigten Programme installiert und die aktuell benutzten Daten gespeichert werden. Eine zweite Platte im PC (nicht nur eine andere Partition!), auf die alle wichtigen Daten gedoppelt werden, hilft gegen einen Plattencrash. Bei RAID-1-Systemen werden alle Daten parallel auf zwei Festplatten gespeichert. Das schützt vor Datenverlust durch eine kaputte Festplatte, aber nicht vor Brand oder Diebstahl.

11.3.2.2. Anschluss über ein Netzwerk:

Die zweite Platte zur Datensicherung ist vom PC entfernt über das Netzwerk verbunden. Sollte der ganze PC zerstört oder gestohlen werden, sind die wertvollen Daten noch auf dem Server vorhanden. Manche Anbieter bieten ihren Kunden zur Datensicherung Platz auf ihren Servern auf einer virtuellen „Wolke“ ***(Cloud)*** an; die Dateien sind dann weltweit verfügbar. Damit überlassen Sie Datenschutz und Datensicherheit aber dem jeweiligen Anbieter, von dessen Gunst Sie dann abhängig sind. Wenn Sie den Vertrag kündigen, haben Sie auch keinen Zugriff mehr auf ihre Daten. Wenn der Anbieter pleite geht oder den Dienst einstellt, sind Ihre Daten verloren oder er wird von einer anderen Firma übernommen, die dann auch Zugriff auf Ihre Daten hat.
Für den Heimbereich gibt es ***NAS*** (Network Attached Storage), kleine Gehäuse mit einer oder mehreren Festplatten (je nach Gerät auch im RAID-Verbund), die direkt an ein Netz angeschlossen werden und auf die jeder zugreifen kann. Hier lassen sich auch Fotos oder die mp3-Sammlung archivieren; wichtig ist die sorgfältige Konfiguration, damit nicht die ganze Welt Zugriff hat und Sie ungewollt zum Anbieter für illegale Downloads werden. In großen Betrieben werden die lokalen PCs mehr und mehr zu Terminals degradiert. Die Rechenarbeit leistet der Prozessor im lokalen PC, die Daten und zunehmend auch die Programme sind auf zentralen Servern abgelegt.
Der Anschluss kann über Kabel (*LAN, Local Area Network*) oder drahtlos (wireless) per *WLAN* erfolgen. Die Kabellösung ist schneller und sicherer. WLAN ist natürlich bequemer, da keine Kabel verlegt werden müssen und die Zahl der Nutzer nicht durch die Anschlüsse am Verteiler beschränkt ist. Es ist aber auch langsamer und störanfälliger, außerdem muss es nach außen gut abgesichert werden.

11.3.2.3. Externe Geräte

Externe Geräte werden nur bei Bedarf an den PC angeschlossen, sonst liegen sie gesichert im feuerfesten Panzerschrank (oder auch nur in der Hosentasche). Einerseits sind sie gut geeignet, um Daten zu transportieren, andererseits auch zur Datensicherung. Ein Laufwerk, das keinen Kontakt zu einem Rechner hat, kann weder ausgespäht noch beschrieben werden und ist somit sicher vor Trojanern, Viren & Co. (Sie können aber infiziert werden wenn sie angeschlossen werden und dann ihre bösartige Fracht auf andere Rechner verbreiten. Es nützt nicht viel, wenn ein Netzwerk mit Firewall und Proxy gegen Bedrohungen von außen geschützt wird und dann von innen per USB-Stick infiziert wird.)

USB steht für universeller seriellerer Bus und trägt das Prädikat universell wirklich zu Recht. USB ersetzt 4 alte Anschlüsse (parallel Port für Drucker, serial Port für Modems, Gameport für Joysticks und die PS2-Anschlüsse für Maus und Tastatur). Dabei ist nicht nur die Geschwindigkeit, sondern auch der Komfort gestiegen. USB-Geräte können bei laufendem PC angeschlossen werden. Außerdem kann USB angeschlossene Geräte mit Strom versorgen. Die abgegebene Leistung ist gering, reicht aber für Sticks, Speicherkarten und kleine Festplatten. Für den Anschluss am PC benutzt man den Steckertyp A (flach und eckig), für Anschlüsse an den Geräten dagegen den Typ B (schmaler, höher, oben abgerundet). Da diese Stecker etwas klobig sind, haben verschiedene Hersteller für kleine Geräte (Kameras, Navis, Tablets) inoffizielle Anschlüsse (***Mini-USB***, Typ B) in ihren Geräten verbaut (mitgeliefertes Kabel gut aufheben, Ersatz ist im Laden manchmal schwer zu bekommen). Um die Verwirrung mit den unterschiedlichen Mini-USB-Anschlüssen zu beenden, haben sich die Hersteller auf ***Micro-USB*** Typ B geeinigt.

USB 1 ist ein nicht benutzter Prototyp. USB 1.1 kann mit vielen Geräten umgehen, erlaubt aber nur eine langsame Datenübertragung (bis 12 MBit/s in *Full Speed*). USB 2 schafft eine Übertragungsrate bis zu 480 MBit/s (***Hi-Speed***) und ist ansonsten kompatibel mit USB 1.1. USB 3 soll Daten mit bis zu 5 Gigabit pro Sekunde (***Superspeed***) zehnmal so schnell transportieren wie USB 2. Die neuen Kabel haben zwei zusätzliche Anschlusspaare, ohne die keine schnelle Übertragung möglich ist. Nur die Stecker vom Typ A (am PC) sind kompatibel. USB-3-Kabel können nicht an USB-2-Geräte angeschlossen werden. Es ist aber möglich, USB-3-Geräte mit USB-3-Kabel an USB-2-Schnittstellen am PC anzustecken, ebenso können USB-2-Geräte mit USB-2-Kabel mit USB-3-Schnittstellen verbunden werden (natürlich nur mit USB-2-Geschwindigkeit). USB-3-Anschlüsse vom Typ A sind meistens innen blau markiert.
Für die USB-Norm gibt es ein spezielles Markenzeichen, einen blauen Stecker mit der Schrift *CERTIFIED USB* (das CERTIFIED steht sehr klein links auf dem Kabel, das B ragt rechts über den Stecker). Ein zusätzliches, oben über den Stecker gezogenes rotes Fähnchen mit der Aufschrift „HI-SPEED“ garantiert die volle Geschwindigkeit für USB 2. Auch für *Superspeed* gibt es ein geschütztes Logo mit einem blauroten Doppelpfeil. Nur bei diesem geschützten Logo können Sie darauf vertrauen, dass Sie die volle Leistung bekommen. Und dieses Logo bezieht sich nur auf die Übertragung, es sagt nichts über die Geschwindigkeit einer Festplatte oder eines Speicherchips aus. Schließlich sollten alle verwendeten Kabel und Hubs (Verteiler) das entsprechende Logo tragen, eine Schwachstelle verlangsamt das ganze System.
Bereits geplant ist USB 3.1 mit verdoppelter Geschwindigkeit und neuen nicht kompatiblen Anschlusssteckern (Typ C), die dann sowohl am Gerät als auch am PC (Host) verwendet werden können und alle bisherigen Stecker ersetzen sollen.

FireWire (korrekt muss es **IEEE 1394** heißen) ist ähnlich aufgebaut wie USB. Es ist älter als USB und wurde zuerst auf Apple-Computern (Mac) eingesetzt. Und für diese ist der Name FireWire gesetzlich geschützt, auf PCs darf es nur IEEE 1394 geben. Andere Hersteller haben sich deshalb andere Namen ausgedacht (z.B. *i.LINK* bei *Sony*), was die Sache nicht übersichtlicher macht. Als die Camcorder digital wurden (MiniDV) gab es erst das langsame USB 1.1. Deshalb setzten viele Hersteller IEEE 1394 ein, um die Daten vom Camcorder an ein Schnittprogramm auf dem PC zu übertragen. Auch einige externe Festplatten waren zusätzlich mit dieser Schnittstelle ausgerüstet. Windows kann mit dem Anschluss des größten Konkurrenten nicht umgehen. Nur wenige PCs sind mit IEEE 1394 versehen (Aufrüstung über Steckkarten ist immer möglich) und der Benutzer muss in jedem Fall die entsprechenden Treiben installieren.

ESATA: Der interne parallele *ATA*-Standard mit breiten Kabeln für den Anschluss von Festplatten wurde durch die serielle Version *SATA* ersetzt. Diese ist wie USB so ausgelegt, dass die Platten im laufenden Betrieb ausgetauscht werden können (hot plug). Damit bietet sich diese Schnittstelle auch für den externen Anschluss einer Festplatte an, die dann mit der gleichen Geschwindigkeit wie die internen Platten angesprochen wird. Es werden aber andere Kabel und Stecker als bei der internen Verkabelung benutzt (Abschirmung) und die externen Geräte werden auch nicht automatisch mit Strom versorgt.

Bluetooth (IEEE 802, benannt nach dem kommunikationsfreudigen Wikinger Harald Blauzahn) ist eine Funkverbindung für kurze Distanzen und reicht nicht viel weiter als ein Kabel. So können Bilder von einem Handy auf einen Drucker ausgegeben werden, ohne nach dem passenden Kabel zu suchen oder ein Headset lässt sich drahtlos mit einem PC oder Handy verbinden. Trotz der kurzen Reichweite kann auch eine Bluetooth-Verbindung abgehört werden, das gilt besonders in der Öffentlichkeit, etwa in einem Warteraum oder einem Zug. Für die Übertragung großer Datenmengen ist Bluetooth nicht geeignet.

11.4. Was soll gesichert werden?

Von Anfang an sollten 3 Bereiche streng voneinander getrennt werden:
Zum ***Betriebssystem*** gehören Programme und Daten, die den Betrieb des Rechners und der angeschlossenen Geräte wie Maus, Bildschirm oder Drucker ermöglichen, z.B. *Windows*, Gerätetreiber oder Zeichensätze. Unsachgemäße Eingriffe können dazu führen, dass der Rechner abstürzt oder nicht mehr startet. Zugriff hat normalerweise nur der Administrator, nicht der einfache Benutzer.
Programme oder ***Anwendungen*** sind Software, die Sie fertig erworben haben, z.B. SPSS, ein Officepaket oder auch ein Spiel. Damit es ggf. Daten mit anderen Programmen austauschen kann, muss es sich in diversen Systemtabellen eintragen. Diese meist recht umständliche Prozedur wird dem Benutzer vom Installationsprogramm (*Setup*) abgenommen. Entsprechend gibt es normalerweise auch ein Deinstallationsprogramm, das (hoffentlich) auch diese Einträge wieder entfernt. Manuelle Eingriffe in die Programmverzeichnisse führen leicht zu Abstürzen oder einem instabilen System und sollten vermieden werden.
Eigene Dateien sind Software, die Sie selbst erstellt haben. Dazu zählen z.B. geschriebene Texte, Präsentationen und eingetippte Daten, aber natürlich auch selbst erstellte Programme oder Makros wie SPSS-Syntax oder Skripte. Legen Sie alle Ihre Daten in ***einem*** Verzeichnis ab. *Windows* sieht dafür das Verzeichnis ***Eigene Dateien*** vor, das für jeden Benutzer separat angelegt wird. Normale User sehen nur das eigene

Datenverzeichnis, Admins haben Zugriff alle Daten auf dem PC. Hier richten Sie für jedes Projekt ein Unterverzeichnis ein, das Sie je nach Umfang und Bedarf weiter unterteilen.
In der Regel ist es unnötig, die gesamte Festplatte zu sichern. Es ist zwar ärgerlich und mühsam, aber mit den Original-CDs kann man Betriebssystem und Anwenderprogramme jederzeit neu installieren (und bei der Gelegenheit vorher nach den neuesten Treibern und Servicepacks im Internet suchen). Eine dateibasierte Sicherung von System und Anwenderprogrammen wird ohnehin nicht funktionieren. Während des Betriebs sammelt sich - besonders unter *Windows* - ständig Datenmüll an (z.B. Reste von längst deinstallierten Programmen oder von Internet-Sitzungen), den Sie ständig mitsichern. Kommt es dadurch zu einem instabilen System, nützt die Sicherung Ihnen auch nichts mehr, da sie die entsprechenden Fehler (oder gar Viren) selbst enthält. Es kann sinnvoll sein, nach einer sauberen Installation des Systems (eventuell plus Updates und Servicepacks) und der Standardprogramme ein Image der Festplatte anzulegen (auch als *Desaster Recovery* bezeichnet bzw. als *Recovery-CD* mit dem PC geliefert).
Wirklich wichtig ist die Sicherung der selbst erstellten Dateien. Bis z.B. die Daten einer Studie ausgewertet werden können, ist meist viel Zeit und Arbeit nötig. Selbst wenn neben den Fragebögen auch die kleinste Korrektur schriftlich festgehalten wurde (was in der Praxis selten der Fall ist), dauert es lange, bis eine verlorene Datei wieder eingetippt und auf den aktuellen Stand gebracht ist. Die optimale Sicherung ist eine Frage des Aufwandes und nicht zuletzt des Geldes (wobei zunehmender Komfort die Kosten deutlich steigen lässt). In der Regel wird man zur Sicherung der eigenen Daten Kopien auf externen Datenträgern anlegen.

11.5. Strategien

Ein wichtige Frage ist es, welche Daten gesichert werden sollen. Natürlich können Sie jedes Mal alle *Eigenen Dateien* sichern (oder die ganze Platte D, oder wo immer Sie Ihre selbst erstellten Daten ablegen). Wenn Sie aber gerade intensiv an Ihrer Promotion arbeiten, reicht es vielleicht, nur das Unterverzeichnis *Diss* regelmäßig zu kopieren. Und Dateien, die sich schnell erzeugen lassen wie SPSS-Outputs oder selbst erstellte EXE-Dateien sind leicht zu verschmerzen, wenn Sie Daten und Syntaxfile bzw. den Quellencode Ihres Programms haben. Temporäre Dateien, wie sie z.B. beim Umstrukturieren eines Datenfiles entstehen, sollten Sie nicht sichern.
Schließlich gibt es beim Sichern mehrere Methoden:
1. ***Vollsicherung:*** Es werden alle Daten kopiert. *Vorteil:* Wenn Sie eine bestimme Datei suchen, befindet sie sich auf jeden Fall auf dem Datenträger. *Nachteil:* Die Sicherung dauert - je nach Datenmenge - relativ lange.
2. ***Inkrementelle Sicherung:*** Es werden nur die seit der letzten Sicherung veränderten Daten kopiert. Dazu muss das ***Archiv-Flag*** korrekt verwaltet werden. Es gehört zu den Datei-Attributen wie der Schreibschutz oder das System-Flag und wird von *Windows* immer gesetzt, sobald eine Datei angelegt oder überschrieben wird. Die Backup-Programme löschen diese Marke, sobald sie eine Sicherungskopie der Datei erstellen. Folglich wurde von allen Dateien mit gesetztem Archiv-Bit noch keine Sicherung angelegt. Der Nachteil des *inkrementellen Backups* ist, dass man nach einer bestimmten Datei unter Umständen lange suchen muss (auf welchem Datenträger ist jetzt die neueste Version?), und auch bei einem kompletten Restore müssen alle erstellten Sicherungen in der richtigen Reihenfolge eingespielt werden. Ein Kompromiss ist es, wenn das Archiv-Flag nur bei einer Vollsicherung gelöscht wird. Dann werden stets alle nicht in der Vollsicherung enthaltenen Dateien kopiert, es gibt also nur die Vollsicherung und *eine* Ergänzung statt einer Ergänzung der Ergänzung der Ergänzung...
Auch Sicherungsbänder oder -platten können kaputtgehen. Außerdem kann es vorkommen, dass Sie selbst eine wichtige Datei aus Versehen löschen oder überschreiben. Nach der nächsten Vollsicherung ist dann auch das Backup dieser Datei weg. Deshalb ist es ratsam, die Datenträger nicht sofort zu überschreiben, sondern auch ältere Backups aufzuheben. Erstellen Sie zuerst die Sicherung „A“. Beim zweiten Mal nehmen Sie einen neuen Datenträger und erstellen darauf die Sicherung „B“. Erst bei der dritten Sicherung wird „A“ überschrieben, bei der vierten dann wieder „B“ usw. So haben Sie immer die letzte und die vorletzte Sicherung zur Verfügung. Natürlich können Sie auch mehr als 2 Sätze anfertigen. Profis empfehlen das „Großvater-Vater-Sohn-Prinzip“. Dazu benötigen Sie 24 Datenträgersätze, die jeweils eine Sicherung aufnehmen können. Diese beschriften Sie mit Montag bis Sonntag (7 Stück), Woche 1 bis Woche 5 (5) und Januar bis Dezember (12). Täglich erstellen Sie eine Sicherung auf dem entsprechenden Tagesband (Mo-So; eventuell auch nur die geänderten Dateien sichern), am Ende der Woche auf das Wochenband (n-te Woche im Monat), am Monatsende auf das Monatsband. Somit haben Sie für die letzte Woche tägliche Backups, für den letzten Monat wöchentliche und für das letzte Jahr monatliche Sicherungen zur Verfügung. Dieses Verfahren wenden besonders Rechenzentren an, die alle Daten eines Rechnersystems routinemäßig sichern müssen.
Schließlich sollten die Backups keine Verbindung zum PC haben und räumlich vom PC getrennt gelagert werden, damit sie nicht dem gleichen *bösartigen Code* (Virus), Brand oder Dieb zum Opfer fallen wie der PC.

Am Ende können Sie alle Dateien, die zu einer Studie gehören, auf einem Datenträger archivieren (wichtige Daten ggf. auch mehrfach sichern oder auf CD brennen) und von Ihrer Festplatte löschen.
Denken Sie daran, Ihre Arbeit immer gut zu dokumentieren! Notieren Sie genau, welche Schritte Sie unternehmen, welche Daten wo stehen und wie und mit welchen Programmen und Makros sie bearbeitet werden. Dazu eignet sich ein Textprogramm oder auch ein einfacher Texteditor. Diese Protokolle sichern Sie zusammen mit den übrigen Daten, dann finden Sie sich auch noch zurecht, wenn Sie die Daten nach längerer Zeit noch einmal brauchen.

P.S.: Computerfachleute zeichnen sich dadurch aus, dass sie sich - im Gegensatz zu Laien - niemals irren können. Das beweist auch die folgende kleine Zitatensammlung aus einem *Stern* von 1996 :-)

- „*Ich glaube, es gibt einen weltweiten Bedarf an vielleicht fünf Computern*“
  (Thomas Watson, IBM-Chef 1943)
- „*Ich habe das ganze Land bereist und mit allen Experten gesprochen, und ich sage Ihnen: Datenverarbeitung ist ein Modefimmel, der nicht einmal dieses Jahr überstehen wird!*“
  (Der Lektor von Wirtschaftsbüchern beim renommierten Verlag *Prentice Hall*, 1957).
- „*Ganz nett, aber ... wozu soll er gut sein?*“ (IBM-Ingenieur über den Mikroprozessor, 1968).
- „*Es gibt keinen erdenklichen Grund, weshalb jemand einen Computer für zu Hause haben sollte*“
  (Ken Olsen, Gründer und Präsident von *digital equipment*, 1977).

7

A

B

C

D

E

F

G

H

I

W

Z