

5. Anmerkungen zur Datenverarbeitung

In den bisherigen Kapiteln wurde der Weg beschrieben, um im Rahmen von ernährungs-epidemiologischen Studien Informationen zu sammeln. Sind die Daten erfasst, dann folgt der letzte Schritt (Abb.15), in dem versucht wird, die Forschungsfrage zu beantworten. Dazu müssen die Daten aufbereitet und gemäß dem Forschungsmodell ausgewertet und schlussendlich muss die Antwort in einem Bericht dokumentiert werden, der der Wissenschaftsgemeinde, den Geldgebern und den Betroffenen zugänglich gemacht wird. Erst mit der Bekanntgabe des Resultats ist der lange Forschungsweg beendet.

Die Gewinnung, Erfassung und Analyse von Daten in ernährungs-epidemiologischen Studien ist im Prinzip ein kontinuierlicher Prozess (Abb. 54), der vor Beginn der Datenverarbeitung einsetzt. Man wählt bewusst das Untersuchungs-Modell aus (s.Kap.2.2.2.), man legt sich auf Variable fest und entscheidet, wie genau man messen will (s.Kap.2.3.4.). Auch bei der Erstellung von Fragebögen wird durch die Art der Frage-Antwort-Formulierung und der Skalen-Bildung die Datenerfassung gestaltet. Am Beginn der Datenverarbeitung steht die Verschlüsselung, d.h. die erhobenen Informationen (Items) müssen mit vereinbarten Symbolen (z. B. Maß-Einheiten) verglichen und eingeordnet werden; sie werden entsprechend dem Untersuchungsmodell bestimmten Variablen zugeordnet (Abb.33). Diese Zuordnung muss den folgenden Anforderungen genügen:

- Eindeutigkeit,
- Ausschließlichkeit,
- Vollständigkeit

Die entsprechenden Aufgaben werden im folgenden Kapitel erläutert. Da bei ernährungs-epidemiologischen Studien die Ernährung des Menschen im Mittelpunkt der Betrachtungen steht und hier die Verschlüsselung besonders umfangreich und problematisch erscheint, wird diese in einem speziellen Kapitel beschrieben (s.Kap.5.2.). Wenn alle Informationen EDV-gerecht verschlüsselt und erfasst sind, dann sind sie in EDV-Geräten gespeichert und man hat eine Daten-Bank (s.Kap.5.3.), wobei die Daten-Banken für Ernährung wiederum besonders hervorgehoben werden müssen. Erst danach können die Daten im eigentlichen Sinne verarbeitet, d.h. gemäß dem Untersuchungsmodell analysiert werden (s.Kap.5.4.). Schließlich erfolgt die Darstellung der Informationen, z. B. in Form von Forschungsberichten.

Wie bei allen anderen vorhergehenden Kapiteln soll auch hier wiederum betont werden, dass die Darstellung nur exemplarisch erfolgen kann.

5.1. Die Verschlüsselung der erhobenen Informationen und die EDV-gerechte Datenerfassung

Die erhobenen Informationen können in ganz verschiedener Form gewonnen, erfasst und dokumentiert sein (Datenträger); neben Fragen und Antworten (auf Fragebögen), können dies Messergebnisse (z. B. Wiegen der aufgenommenen Nahrung, anthropometrische Maße, Labordaten, usw.) und Abschätzungen (z. B. Erinnerung an das Essen in einem zurückliegenden Zeitraum) sein. Der Daten- bzw. Informationsträger kann ebenso in Form von Bildern (wie z. B. Photographien und Filme) vorliegen.

Das Prinzip der Auswertung besteht darin, die Informationen zu ordnen und zu vergleichen, wobei das Untersuchungsmodell der Leitfaden ist. Heute wird dieses früher mühsame, langweilige und zeitraubende manuelle Verfahren in aller Regel durch EDV ersetzt. Dabei kann der gedankliche Prozess - was wem wie zugeordnet wird - nicht durch den Computer erfolgen, sondern die EDV-Geräte müssen genau "gesagt" bekommen, mit welchen Daten sie

"gefüttert" werden. Wo welche Information "abgelegt" werden soll, damit man sie wieder findet und z. B. auszählen kann. Diese Zuordnungen stellen den Schlüssel dar, der meist auch Kode genannt wird.

Diese Anweisungen an den Computer werden in einem Kode-Plan festgelegt, der von dem EDV-System abhängig ist. Es gibt zur Erfassung von Daten bestimmte Computerprogramme (Software), wie z. B. SPSS (Statistical Package for Social Sciences), das es heute neben der Version für Großrechenanlagen, auch in PC-Versionen gibt.

Viele weitere Auswertungsprogramme sind auf dem Markt (z. B. BMDP, SAS, CLUSTAN, GLIM, LISREL, MDSX, PLOTIT, SIR/DBMS und STATGRAPH); damit können Daten in verschiedener Weise bearbeitet werden, das betrifft statistische Analysen und auch die (graphische) Darstellung von Ergebnissen. Doch zuerst müssen die Daten in die jeweiligen EDV-Anlagen eingegeben werden.

Dies kann in verschiedener Weise geschehen. Die alte Hollorith- Karte wird zwar nicht mehr benutzt, doch sind deren Prinzipien noch erkennbar. So gibt man z. B. bei SPSS die Daten in Zeilen ein, die man meist auf 80 Spalten begrenzt. Das ist im Prinzip heute nicht mehr notwendig, doch es hat praktische Hintergründe, wie die der Übersichtlichkeit und der Platz auf dem Bildschirm. In jede Spalte können alle Ziffern und Zeichen, die einem EDV-System zur Verfügung stehen, eingegeben werden.

In der Anwendung ist es jedoch einfacher, wenn man Ziffern ("numerische Eingabe") benutzt und weitgehend auf Eingabe von Buchstaben und Sonderzeichen ("alpha-numerische Eingabe") verzichtet.

Die Erstellung eines Kodeplanes kann und soll bereits mit den Überlegungen zur Methoden-Auswahl beginnen. Man muss wissen, welche Variablen in welcher Form (Indikatoren, Skalen, Messvorschriften) erfasst werden sollen (Abb.33). Man baut daraus die Erhebungs-Instrumente (z. B. Fragebogen, Ergebnis-Protokoll-Listen) auf. Man hat sich dabei entschieden, wie die Information erfasst wird - z. B. geschlossene Frage (Antwortvorgaben), offene Fragen; wie genau gemessen wird (z. B. Körpergewichtsmessung auf 0,1kg). Die Verschlüsselung sollte soweit wie möglich bereits auf dem Erhebungsbogen festgelegt sein.

Man kann die Variable bzw. den Indikator mit einer Kurzform kennzeichnen ("Label"), wobei es günstig ist sinnvolle, d.h. leicht merkbare Abkürzungen zu wählen, und nicht beliebige Reihenfolgen, VAR1, VAR2,... VARmno, deren Bedeutung man sich schwer merken kann. Schließlich kann man auflisten, warum man welche Variable in das Studien-Design aufgenommen hat (z. B. Merksätze zu den Arbeitshypothesen). Damit entsteht ein grober Kode-Plan in Form eines Variablen- Index.

Die Reihenfolge und Form der Daten, so wie sie den EDV-Systemen eingegeben werden sollen, muss danach in einem formellen Kode-Plan festgelegt werden.

Dieser muss dem Computer mitgeteilt werden, er muss wissen, was wo in welchem Speicher steht; damit er dann später die Daten verarbeiten kann.

In der Regel beginnt man mit der Versuchspersonen-Nummer. Die weitere Reihenfolge ist zwar im Prinzip beliebig, doch werden praktische Gesichtspunkte berücksichtigt, z. B. die Reihenfolge auf dem Frage- bzw. Erfassungsbogen, ähnliche Inhalte, usw. Als nächstes wird der Wertebereich festgelegt; dieser ergibt sich aus den (Vor-)Kenntnissen über die Studie; und damit wird die Zahl der benötigten Spalten pro Variable festgelegt. Bei der Festlegung der Positionen für die einzelnen Spalten sollten auch praktische Gesichtspunkte - vor allem die der Erleichterung der Daten-Eingabe - berücksichtigt werden. So ist es günstig durch Lücken bzw. freie Spalten die einzelnen Variablen zu trennen; die Zeilen nicht zu lang zu machen,

usw. Man kann dann im weiteren aus den einzelnen Variablen noch verschiedene Umrechnungen vornehmen. Als nächstes kann man festlegen, was geschieht, wenn in der betreffenden Position ein Wert fehlt ("missing value").

Die Dateneingabe erfordert sehr viel Konzentration. Mit entsprechender Software kann man dies jedoch erleichtern: man programmiert die Eingabe entsprechend dem jeweiligen Kodeplan so, dass entsprechende Bildschirmmasken erscheinen, die Abfragen - welche Information in die betreffende Position gehört. Doch dies erfordert wiederum Spezialkenntnisse im Umgang mit der EDV.

An dieser Stelle ist es notwendig, sich Gedanken über die Datenmatrix zu machen; dazu werden die verschiedenen Angaben zu den einzelnen Variablen in ihrer Beziehung zu den betreffenden Studienteilnehmern betrachtet. Liegen von jedem Studienteilnehmer alle Angaben vor, dann ergibt sich eine vollständige rechteckige Datenmatrix. Solchen idealen Daten-Anordnungen sind statistische Berechnungen leichter zugänglich, allerdings sind sie bei großen ernährungsepidemiologischen Studien nicht die Realität. Aus den verschiedensten Gründen gibt es Lücken.

Die in die EDV-Systeme aufgenommenen Daten können fehlerhaft sein, man muss sie in einem nächsten Schritt kontrollieren. Das beinhaltet verschiedene Gesichtspunkte, wie Vollständigkeit der Eingabe und Richtigkeit der Daten. Dieser Schritt kann sehr umfangreich sein, ist aber äußerst wichtig. Man kann mit einer reinen Häufigkeitsauszählung der Variablen der Datei beginnen und sich diese ansehen. Dabei fallen leicht offensichtliche Fehler auf, die aus folgenden Ursachen stammen können:

- Fehler im Kode-Plan bzw. Steuerdatei (z. B. zu wenige Records pro Fall, Vergessen eines Kodes, falsche Spaltenangaben, usw.);
- Fehler bei der Daten-Eingabe (Spalten-Verschiebungen, falsche Zeichen, Tippfehler, Verwechslung, usw.).

Eine Reihe von Fehlern können so offensichtlich sein, dass sie gleich zu finden sind. Andere können durch entsprechende Plausibilitäts- Programme gesucht werden; in der einfachsten Form sind es Angaben bzw. Kontrolle der möglichen Wertebereiche. So ist z. B. eine Erythrozytenzahl von >20 Terra/Liter unglaublich; oder ein Alter >80 Jahre, wenn bekannt ist, dass der älteste Teilnehmer 72 war, usw. Man kann nach unmöglichen Kombinationen suchen, wie z. B. schwangere Männer oder Verheiratete, die jünger als 10 Jahre sind. Für diese Art der Prüfung sind spezielle EDV-Programme zu erstellen.

Daten-Kontrollen können auch "einfach" dadurch erfolgen, dass eine Stichprobe von Datensätzen mit den Ursprungsdaten verglichen werden, also Korrektur gelesen wird. Eine besonders aufwendige, aber wirkungsvolle Kontrolle ist die unabhängig voneinander ablaufende, doppelte Verschlüsselung und Eingabe der Daten. Alle Abweichungen werden dann automatisch mit entsprechenden EDV-Programmen identifiziert.

Fehler müssen mit Hilfe der Originaldaten überprüft werden. So ist es wichtig, dass dies schnell geschieht - denn wenn nicht mehr Rückfragen bei denen möglich sind, die die Daten erfasst haben, dann können die Unstimmigkeiten auch weniger gut bereinigt werden. Schließlich müssen an dieser Stelle Kriterien festgelegt werden, die zum Ausschluss einzelner Daten bzw. ganzer Datensätze führen. Dieser lange Prozess, der schließlich zu einer Datenbank führt, mit der man dann seine gewünschte Daten-Verarbeitung vornehmen kann, ist bei Ernährungsdaten voller spezifischer Tücken, so dass dies im folgenden Kapitel besonders beschrieben wird.

5.2. Die Verschlüsselung von Ernährungsdaten

Die Daten zur Nahrungsaufnahme bzw. zum Ernährungsverhalten werden überwiegend in offener Frageform erfasst. Man fragt die Studienteilnehmer, was sie gegessen haben bzw. man protokolliert den Verzehr (s. Kap. 3.1.) und verschlüsselt anschließend die Angaben. Der Hauptgrund für diese Vorgehensweise ist in der großen Vielfalt von Lebensmitteln zu suchen. Allerdings wird der einzelne Mensch bei weitem nicht alles essen, nicht das mögliche Potential ausschöpfen. So verzeichnen Ernährungserhebungen aus unseren Lebensräumen Größenordnungen von 100-300 verschiedenen Angaben pro Studienteilnehmer, bei Studien in Entwicklungsländern kann dies um einiges geringer sein (10-30). Trotzdem müsste man, wollte man von vornherein alle Lebensmittel, die im Verlaufe der Erhebung angegeben werden könnten, verschlüsseln, eine umfangreiche Liste vorbereiten. Nur in den speziellen Fällen, bei denen man sich aus den verschiedensten Gründen nur auf wenige ausgewählte Lebensmittel(gruppen) beschränkt, können geschlossene Erhebungsbögen benutzt werden (s. z. B. Nahrungsmittel- Frequenzen und Checklist-Protokoll).

An sich ist die Verschlüsselung der erfassten Ernährungsangaben nicht schwierig; den einzelnen Lebensmitteln werden bestimmte Nummern zugeordnet. Diese Schlüsselzahlen bzw. Kode-Nummern könnten nach vielen verschiedenen Prinzipien gewählt werden, z. B. auch nach einer alphabetischen Reihenfolge der Namen (z. B. Aal = 1 und Zwiebelsuppe = 9999). Eine andere Form wäre die direkte Benutzung der EAN- Strichkodierung.

Es ist jedoch vernünftiger, das Ziel der Studie als Richtschnur zu verwenden und das ist in den meisten Fällen die ernährungsphysiologische Bewertung der Nahrungsaufnahme, das schließt die Ordnung der Nahrungsmittel in bestimmte ernährungsphysiologisch gleichwertige Gruppen ein (Fleisch, Obst, Gemüse, usw.). Man will dabei die einzelnen Aufnahmen an Nährstoffen bzw. sonstigen Nahrungsmittel-inhaltsstoffen erfassen. Je nach dem Auswertungsziel werden verschiedene Lebensmittelgruppierungen vorgenommen; kommt es nur auf den Energiegehalt der Nahrung an - so genügt bei energiearmen Lebensmitteln (wie Obst, Gemüse usw.) eine grobe Klassifizierung; will man jedoch die verschiedenen Fettsäuren erfassen, dann müssen die Angaben über die einzelnen Öl- und Fettsorten möglichst detailliert verschlüsselt werden; und wieder anders sieht es aus, wenn die Aufnahme an verschiedenen Vitaminen, Mineralstoffen oder Spurenelementen beurteilt werden soll. Diese traditionelle, naturwissenschaftlich ausgerichtete Verschlüsselung soll auch hier im Mittelpunkt der Betrachtungen stehen, ohne andere, ebenfalls mögliche wichtige Aspekte der Auswertung und damit entsprechenden Verschlüsselung der Ernährungsdaten unterschlagen zu wollen.

Die Nahrungsaufnahme stellt keineswegs nur einen Akt der Nährstoffaufnahme dar, sondern sie ist auch ein Ausdruck des Ernährungsverhaltens. Folglich können Ernährungserhebungsdaten auch solche Informationen beinhalten, dazu liegen jedoch kaum Auswertungsverfahren vor. Solche Analysen werden jedoch möglich, wenn mit der jeweiligen Nahrungsaufnahme folgende Informationen in Beziehung gebracht werden:

- Art der Mahlzeit,
- Zeitpunkt und -dauer,
- Verzehrort und -situation,
- Reihenfolge des Verzehrs, Kombinationen, Zubereitungen, Menüs.

Die verzehrten Lebensmittel können neben ernährungsphysiologischen Besonderheiten, auch andere Charakteristika aufweisen (Abb.56), die für das Untersuchungsmodell als wichtig erachtet werden, wie:

- industriell erzeugte Lebensmittel, Marken-Name, Discount- und Convenienceprodukte, Fertiggerichte, Fast-Food, usw.;
- selbst und alternativ erzeugte Produkte;
- regionale Herkunft; landsmannschaftliche, regionale Küche, importierte Ware (Herkunftsland; Dritte-Welt; Boykott-Ware, usw.);
- Gesichtspunkte der Sinneswahrnehmung (Farbe, Geschmack, Konsistenz (knusprig, breiig, usw.), Geruch, usw.);
- ernährungsökologische Gesichtspunkte (Energie- und Rohstoff- Verbrauch bei der Herstellung, Verpackung, usw.);
- ernährungsökonomische Gesichtspunkte (Nahrungspreise).

Dies muss selbstverständlich durch die entsprechende Form der Erhebung berücksichtigt werden.

Man erkennt, dass der Lebensmittel-Schlüssel allein durch diese Gesichtspunkte sehr vielfältig sein kann; es kommen darüber hinaus noch weitere Gesichtspunkte hinzu. Hier sei an die stoffliche Dimension der Ernährung erinnert; Lebensmittel sind biologischer Herkunft, d.h. es sind keine "festen Stoffmischungen", sondern ihre Zusammensetzung hängt von den Produktionsbedingungen - wie Boden- und Futterqualität - ab; sie ändern sich im Verlauf des Reifeprozesses, nach der Ernte, bei der Verarbeitung, bei der Lagerung, bei der Zubereitung. Die aufgenommene Nahrung unterliegt Verdauungs-Prozessen, hierbei spielen Nahrungsmittel-Interaktionen eine Rolle. Lebensmittel werden im allgemeinen nicht isoliert verzehrt, sondern in bestimmten Zusammenstellungen (Rezept, Gericht, Menü), die ganz individuell sein können (und teilweise "Küchengeheimnisse" sind; aber auch "Industriepatent" - z. B. das "Geheimnis der Coca Cola-Rezeptur"). Schließlich muss man berücksichtigen, dass manche Lebensmittel ganz verschiedene Namen haben, obwohl sie fast gleich sind; umgekehrt kann sich unter einem gleichen Namen sehr Verschiedenes verbergen ("Gulasch", "Hamburger", usw.).

Neben der Identifizierung der einzelnen erfassten Lebensmittel gehört zur ernährungsphysiologischen Berechnung der Nahrungsaufnahme die Mengenangabe dazu. Diese kann in verschiedenen Weisen erfasst sein (Wiegen, Messung mit Standard-Maßen, Abschätzen mit "Modellen", bloße Erinnerungs-Abschätzung, Angaben in Portionen, Verpackungsaufschriften, usw.).

Von den jeweiligen Lebensmitteln müssen Informationen über ihre Zusammensetzung vorhanden sein. Übliche Nährwert-Tabellen enthalten ca. 1000-2000 Lebensmittel, also wesentlich weniger als es Lebensmittel gibt. Durch Abschätzen oder durch aufwendige Analysen müssen Inhaltsstoff-Lücken ausgeglichen werden. Durch einfache Multiplikationen zwischen den betreffenden Konzentrationen in den einzelnen Lebensmitteln mit deren erfassten Mengen erhält man die (Nähr-)Stoffaufnahmen; würde eine nicht bekannte Information mit = 0 angesetzt, so würde *keine Aufnahme* als Ergebnis resultieren, was eine falsche Berechnung darstellt.

Aus den vorgenannten Überlegungen ergibt sich zwangsläufig, dass die Verschlüsselung der Ernährungsdaten nicht total standardisiert werden kann. Man kann nicht ungeprüft auf einen "Food-Kode" zurückgreifen; es gibt keinen fertigen und vollständigen Lebensmittel-Schlüssel. Man muss vielmehr das Bekannte mit den eigenen Erfordernisse der jeweiligen Untersuchungssituation vergleichen. Diesen individuellen Eigenheiten stehen die Probleme der Vergleichbarkeit der Daten zwischen verschiedenen ernährungsepidemiologischen Studien gegenüber, denn dazu ist ein einheitlicher Lebensmittel-Schlüssel wichtig; doch diese

Standardisierungsbemühungen können nur von Institutionen (und nicht von einzelnen Forschern) übernommen werden. Das liegt in der Natur der Sache begründet, denn niemals wird ein Lebensmittelschlüssel fertig sein, er muss immer wieder auf den neuesten Stand gebracht werden ("up-date"). Das geschieht z. B. durch Beobachtung des Marktes, der Ernährungsgewohnheiten, also in Rückkopplung mit entsprechenden empirischen Erhebungen. Der riesige Umfang von solchen Informationsfeldern kann nicht von einer kleinen Gruppe abgedeckt werden, denn solch ein "Kode-Buch" ist ein sehr umfangreiches Werk. Es ist jedoch notwendig, Richtlinien zum gleichen Vorgehen bei der Verschlüsselung zu vereinbaren, und den Rahmen für entsprechende Aufgaben festzulegen. Dazu nun einige Hinweise.

Bei der Konzeption eines Lebensmittel-Schlüssels müssen hinreichende Informationen zu folgende Bereiche vorliegen:

- Welche Arten von Lebensmittel sind bei den geplanten ernährungs-epidemiologischen Studien zu erwarten? Welche Arten bzw. welche Differenzierungen erscheinen aufgrund des Untersuchungszieles als wichtig, welche können vernachlässigt werden?
- Von welchen Lebensmitteln existieren welche Angaben in welcher Qualität über welche Inhaltstoffe? Welche Nährwert-Tabellen stehen zur Verfügung?

Lebensmittel-Kodes werden häufig nach ernährungsphysiologischen bzw. biologischen Gesichtspunkten geordnet; sie sind demnach ähnlich wie Nährwert-Tabellen aufgebaut. Weitere Untergliederungen betreffen die Nährwertgehalte (z. B. Fettgehaltsstufen) und Verarbeitungsstufen (ganzes Nahrungsmittel, Teile davon, roh, gegart, in Dosen, getrocknet, usw.)

Es ist nützlich, daraus hierarchisch gegliederte Zahlen-Kodes zu bilden. Aus 10 Nahrungsmittel-Obergruppen entstehen jeweils mit jeder weiteren Stelle 10 weitere Untergruppen. Ein in England häufig benutzter Kode kommt zwar mit drei Stellen aus (000-999) und damit können 1000 Lebensmittel verschlüsselt werden (Dunn Nutrition Laboratory, Cambridge); solch ein Kode ist jedoch sehr eng, hat kaum Lücken für neue Lebensmittel; so werden 4 Stellen als günstiger angesehen. Hier können bereits viele verschiedene Gesichtspunkte hierarchisch berücksichtigt werden (Abb.54). Bedingt durch die rasche EDV-Entwicklung, können die Kodes noch großzügiger angelegt werden. Bei der Kodierung der Erhebungen des US Department of Agriculture werden 3+4 Stellen benutzt, wobei die ersten 3 Stellen die üblichen Bereich der Lebensmittel umfassen und die weiteren 4 deren spezifische Eigenarten.

Abb.54 Der Nahrungsmittel-Kode GLANZ (Gießener Liste aller Nahrungsmittel und -Zubereitungen (Bodenstedt, Oltersdorf et al., 1983)

Dieses Prinzip der Zusatz-Kodes erscheint als zukunftsweisend, denn damit kann über den "Kern-Kode" einerseits die Vergleichbarkeit hergestellt werden und durch "Zusatz-Kodes" wird auf individuelle Belange der jeweiligen Untersuchungsgruppe eingegangen. Die zusätzlichen Spalten des Kodes können mit ganz verschiedenen Übereinkünften versehen werden. Diese Gedanken leiteten auch die Aufstellung des Kodes, der in der EMSIG- Studie entwickelt und benutzt wurde. Die Gießener Liste aller Nahrungsmittel und Zubereitungen (GLANZ) hat einen 4stelligen "Kern-Kode", deren Ordnungssystem auf biologischen und lebensmitteltechnologischen Gesichtspunkten basiert. Da wir einen großen Teil unserer

Nahrung in (fremd-)zubereiteter Form aufnehmen - wobei die einzelnen Zutaten bzw. die verwendeten Lebensmittel den Studienteilnehmern häufig unbekannt sind; z. B. industriell Vorgefertigtes (Fertiggerichte, Convenience Food), Außer-Haus-Verzehr mit kommerziellen (Kantinen, Restaurants) und privaten Charakter (Einladungen, Feste, usw.) - findet man in Ernährungserhebungen nicht nur Lebensmittel, sondern auch Zubereitungen. Um diese Informationen auswerten zu können, muss man auch diese Angaben aufschlüsseln. So wurde neben der Liste der Lebensmittel auch eine Liste von Zubereitungen angelegt. Als Grundlage dienen Standard-Rezepte. Die Angaben für die Zubereitungen wurden darüber hinaus mit Rezeptangaben durch die Studienteilnehmer bzw. durch Nachfragen bei Außer-Haus-Verpflegungseinrichtungen fortlaufend ergänzt. Die Angaben bei den Zubereitungen enthalten die Namen der Lebensmittel und damit sind beide Ebenen in Verbindung zu bringen.

Neben dem vierstelligen Kode für Lebensmittel und Zubereitungen wird auch die Originalbezeichnung der Studienteilnehmer aufgenommen, da der Namen eine Reihe von weitergehenden Informationen enthalten kann, wie z. B. regionale Herkunft, Markenzugehörigkeit, persönliche Note, usw. Daneben wird der Listen-Name eingegeben, der der entsprechenden Lebensmittel- bzw. Zubereitungs-Kode-Nummer entspricht. Die Zuordnungen zwischen den verschiedenen Namen und den Kode-Nummern müssen vereinbart und festgelegt werden. Durch Sonderkennzeichen - wie Kode-Buchstaben - wird die Identifizierung von Eigenrezepten und den dazugehörigen Lebensmitteln ermöglicht (z. B. R = Name eines Eigenrezeptes; L = Lebensmittel aus einem Eigenrezept; N = Name aus Standard-Rezept-Liste).

Der numerische Kode wird durch eine Reihe weiterer Dimensionen ergänzt. So wird die Kombination bzw. die Zusammengehörigkeit der verschiedenen Lebensmittel und Zubereitungen bei einer Mahlzeit festgehalten. Mit der Kombinationsvariable 1 wird die Mahlzeit aufgegliedert. Dabei gibt es eine zentrale Kombination (z. B. Fleisch und Beilage = 2); alles was davor genannt wurde erhält den Kode = 1, und die nachfolgenden Angaben = 3., während Getränke = 4 sind. Mahlzeiten bei denen kein zentrales Lebensmittel zu erkennen ist werden durchgehend mit 2 kodiert. Bei der Kombinationsvariable 2 werden zusammengehörende Lebensmittel mit einer gemeinsamen Ziffer versehen, wie z. B. Milch und Kaffee; Brot und der Belag; Bestandteile der Rezepturen der Zubereitungen, usw. In einer weiteren Spalte wird die Nennung des Markennamens verschlüsselt; analog gibt es Kodes für die Zubereitungsart (Garmethode).

Ein wichtiger Bereich ist die Verschlüsselung der Mengenangaben; sie erfolgt im Prinzip in Ziffern und der Dimension Gramm. Doch eine in die Datei aufgenommene Angabe, kann mit der gleichen Bezeichnung von ganz unterschiedlicher Datenqualität sein. Die Menge kann abgewogen sein (Kode = 1), es können (haushaltsübliche) Portionsangaben verwendet sein (Kode = 2); die Angaben können unüblich, aber spezifiziert sein (Kode = 3), doch sie können auch unspezifisch und ungenau sein (Kode = 4). Für alle diese Fälle müssen bei den jeweiligen Lebensmitteln bzw. Zubereitungen Vereinbarungen zum Verschlüsseln der Menge getroffen und festgehalten werden.

Die Beziehung zwischen der aufgenommenen Nahrung und der Verzehrsituationen kann relativ einfach hergestellt und verschlüsselt werden. Die Angaben im Kopf des EMSIG-Ernährungsprotokoll-Bogens sind bereits vorkodiert bzw. lassen sich direkt übernehmen, das betrifft Datum und Dauer. Offene Kategorien - wie z. B. Angabe der Mahlzeit - werden jeweils analog zu den bereits festgelegten Kategorien verschlüsselt.

In Fällen bei denen nicht der einzelne Studienteilnehmer die Erfassungseinheit ist, sondern z. B. der Haushalt, muss die Anwesenheit der verschiedenen Essens-Teilnehmer verschlüsselt werden.

Die einzelnen Eingaben sind in einem bestimmten Eingabe-Schema festgelegt. Nach diesem Prinzip der Verschlüsselung der Ernährungsinformationen, kann der "Kern-Kode" je nach Bedarf des Untersuchers mit einem engeren oder weiteren Rahmen ergänzt werden. Dieses GLANZ-Schema ist in verschiedenen anderen Lebensmittelschlüsseln wiederzufinden (*Arab* 1988, *Häußler* et al. 1990, *Petot* 1987).

Es wird deutlich, dass für die Verschlüsselung der Ernährungsdaten auch bei einer weniger aufwendigen Form der Kodierung - wie es für GLANZ beschrieben wurde - eine Vielzahl von Informationen gesammelt werden müssen, und dass viele Übereinkünfte festzulegen sind. All diese müssen dokumentiert werden - es entstehen umfangreiche Kode-Bücher, die aus einzelnen Dokumentationsblättern bestehen. Darauf werden die verschiedenen Informationen festgehalten. Für jedes Lebensmittel müssen entsprechende (Nähr-)Stoffangaben verfügbar sein; man kann hier neben Nährstoffen viele weitere Stoffe mit aufnehmen, eben solche Stoffe, die in den möglichen Untersuchungsmodellen als interessante Variablen eingehen. Man darf spezielle "Arzneien" und "Stärkungsmittel" nicht vergessen, die gerade bei der Zufuhr von Vitaminen und Mineralstoffen bzw. Spurenelementen einen individuell großen Beitrag leisten können.

Die Herkunft der Informationen muss festgehalten werden, manchmal muss die Information einfach in einen plausiblen Rahmen gestellt werden bzw. man muss den Bedarf nach entsprechenden Informationen, z. B. Notwendigkeit von chemischen Analysen herausstellen. Die fehlenden Angaben ("missing data") sind ein großes Problem; würde man alle Unsicherheiten weglassen, also gleich Null setzen, dann wäre das der "größte anzunehmende Fehler". Hat man z. B. die Information "in Fett gebraten", dann kann man nicht die Fettzufuhr = 0 ansetzen, nur weil man nicht weiß, welches von den vielen verschiedenen Fetten es war. Analog ist "1 Scheibe Wurst" eine wohl unsichere Information, doch es war nicht = 0. Gibt es für das betreffende Lebensmittel noch keinen Wert in der Inhaltstabelle, so kann dort auch nicht unbegründet = 0 stehen; man muss abschätzen - z. B. aufgrund von ernährungswissenschaftlichen Kenntnissen und Vergleichen mit ähnlichen Lebensmitteln - wie ein wahrscheinlicher Wert lauten könnte. Bei der Verschlüsselung kann dann aber der Eindruck entstehen, die Angabe wäre zuverlässig. So muss man gegebenenfalls die Qualität der Information mit verschlüsseln.

Ebenso müssen die Übereinkünfte festgelegt werden, die durch ungenaue Erhebungsangaben notwendig werden - das betrifft alle Bereiche: den Namen des Lebensmittels bzw. der Zubereitung, die Portionsgrößen, die Mengenangaben, die Benennungen des Garverfahrens, die Kombinationen, der Zustand (roh bzw. verzehrfertig), usw.

Die Portionsgrößen (wie viel g ist 1 Teelöffel Zucker; 1 Ei, 1 Brötchen, usw.) können (analog wie bei Rezepten) "Standard-Büchern" entnommen sein. Man weiß allerdings zu wenig darüber, inwieweit sich die verschiedenen Bevölkerungsgruppen bzw. die einzelnen Studienteilnehmer daran orientieren. So können Gewicht von und Vorstellungen über "1 Scheibe Wurst" von der Wurstsorte abhängen, aber auch nach Region und Gesellschaftsschicht sehr stark variieren. Auch die Anwesenheit von Essens-Gästen hat einen Einfluss darauf, wie dick die Wurtscheibe ausfällt. Löffel- und Tassen-Größen sind der Besteckmode unterworfen; Portionsgrößen werden von der Lebensmittelindustrie auf die Zielgruppe abgestimmt oder hängen von der Kalkulation einer Kantine oder eines Gastwirtes ab. Folglich sind auch Portionsgrößen für die einzelnen Studien individuell zu vereinbaren.

Letzten Endes muss heute eine solche Informations-Sammlung zu einer Ernährungs-Fakten-Datenbank führen, denn diese Informationssammlung wird nie beendet sein - das Lebensmittelangebot ist genauso lebendig, wie der Forscherdrang, der immer wieder neue Informationen zu Lebensmitteln schafft. Solche EDV-gespeicherten Informationen können dann auch schnell notwendige Hilfsmittel zur Kodierung der Erhebungs-Daten liefern - wie vollständige umfangreiche Listen des Nummern-Kodes der Lebensmittel und Zubereitungen; die Synonyme können alphabetisch sortiert sein; es sind Listen für die jeweiligen Garverfahren zu erstellen, usw.

Der nächste Schritt bei der Verschlüsselung der Ernährungsdaten stellt die eigentliche Eingabe in das EDV-System dar und dann gilt es die gespeicherten Daten zu überprüfen. Bei dem aufwendigen und langwierigen Prozess des Überführens der Daten von den Erfassungsbelegen (z. B. Ernährungsprotokollen) zu der eigentlichen Eingabe können viele Fehler geschehen. Man kann sich einfach bei der Eingabe vertippen - falsche Kode-Zahl, falsche Menge, falsche Kommastelle, usw. Man kann bestimmte Angaben im Erhebungsbogen "überspringen", z. B. ganze Zeilen vergessen. Die Prüfung kann nach den gleichen Prinzipien erfolgen, die sie für die anderen Daten schon genannt wurden. So wird man sich die Eingabe erst einmal einfach ausdrucken lassen, manches (wie z. B. Spalten-Verschiebungen) wird dann schnell "offensichtlich".

Eine aufwendige Kontrolle besteht darin, dass die Verschlüsselung parallel erfolgt und Unstimmigkeiten vom EDV-System zur besonderen Kontrolle aussortiert werden. Analog können entsprechende EDV-Programme geschrieben werden, die merkwürdige Angaben auswerfen - z. B. bestimmte hohe Verzehrangaben (z. B. auch je nach Personeneigenschaften, z. B. für Männer, Kinder). Daneben kann das Prüfprogramm weitere Plausibilitäts-Merkmale bzw. -Beziehungen beinhalten. Im Programm GLANZ ist ein entsprechendes Programm PRUEF eingebaut.

Erst nach entsprechender Korrektur der eingegeben Ernährungsdaten hat man eine Datenbank mit der dann die Auswertungen vorgenommen werden.

Zur Nahrungs-(Energie)-Bilanz gehören auch Angaben zum Bedarf, die teilweise relativ leicht zugänglich und zu verschlüsseln sind, wie z. B. Alter, Geschlecht, Größe und Gewicht. Dazu kommt als weiterer umfangreicher Informationsbereich, die Angaben zur körperlichen Aktivität. Die Methoden dieses Variablenbereiches (s. Kap. 3.2.) sind ähnlich vielfältig und die Informationen ähnlich umfassend, wie die der Erfassung der Nahrungsaufnahme. So erfordert auch deren Verschlüsselung und Eingabe ähnliche Überlegungen und Aufwand. Dazu ist anzumerken, dass der Bereich der körperlichen Aktivität bisher in ernährungsepidemiologischen Studien unzureichend berücksichtigt wurde, so gibt es, verglichen mit den Ernährungsdaten, viel weniger Auswertungserfahrung; einschließlich der Verschlüsselung der Aktivitätsdaten.

Die Informationen, die durch die Befragung bzw. die Protokollierung des Aktivitätsverhaltens gewonnen werden, können sowohl ernährungsphysiologisch - vor allem die Berechnung des Nahrungsenergie-Bedarfs - aber auch sozialwissenschaftlich ausgewertet werden. Allerdings erfordert beides eine unterschiedliche Kodierung der Informationen.

Bei der ernährungsphysiologischen Auswertung wird der Energiebedarf aus Angaben zum Grundumsatz und zum Leistungsumsatz berechnet. Dazu werden Informationen über die Körpermasse des Menschen, das Alter, Geschlecht, usw. benötigt und die Angaben über die verschiedenen Tätigkeiten der Menschen müssen nach Intensitäts-Stufen eingeteilt werden. Danach kann die Verschlüsselung erfolgen. Man kann dann durch Multiplikation der jeweiligen Intensitäts-Stufen mit der entsprechenden Zeitdauer den Arbeitsumsatz berechnen.

Bei einer solchen Verschlüsselung gehen jedoch viele Informationen verloren; alle Aktivitäten "gerinnen" zu einigen wenigen Intensitätsstufen. Man kann, analog wie bei Lebensmitteln, die verschiedenen Aktivitäten mit umfangreicheren, hierarchisch gegliederten Ziffernfolgen versehen. Eine entsprechende Auswertungsroutine gibt es jedoch noch nicht, sondern höchstens verschiedene isolierte Ansätze. In Analogie zu der "flexiblen" und offenen Verschlüsselung bei Ernährungsdaten (s.o. GLANZ) wurde ein Auswertungs-Verfahren entwickelt, das neben dem zentralen Kode für die Tätigkeiten nach den Intensitätsstufen, versucht möglichst viele Informationen aus den entsprechenden Tätigkeits-Protokollen zu erfassen. Im GAST (Gießener Auswertungs-Schema für Tätigkeitsprotokolle) werden deshalb die einzelnen erfassten Aktivitäten der Studienteilnehmer in direkter, alpha-numerischer Weise eingegeben. Dies wird ergänzt durch die Informationen über den Erhebungstag, die Zeit, die Zeitdauer und der Charakterisierung des Tages.

Die verschiedenen Tätigkeitsangaben sind relativ leicht zu dokumentieren, hier genügt im Gegensatz zum viel umfangreicheren Ernährungs-Kode-Buch eine Art lexikalisch alphabetisches Verzeichnis. Problematisch ist jedoch die richtige Zuordnung der verschiedenen Tätigkeiten zu den jeweilig festgelegten Intensitätsstufen, hier gibt es große intra- und interindividuelle Streubreiten. Fast jede Tätigkeit kann mit geringer, aber auch mit hoher Intensität ausgeführt werden, und es sind sich überlagernde Tätigkeiten möglich (z. B. Zeitung beim Frühstück lesen).

Selbstverständlich müssen auch diese Eingaben überprüft werden, dazu werden prinzipiell dieselben Verfahren angewendet, wie sie oben bei den Ernährungsdaten angeführt wurden. Man sieht sich die Ausdrücke der Eingabe-Datei an, man entwickelt Plausibilitäts-Prüfprogramme, man liest Korrektur, usw. Erst dann hat man eine Datenbank.