PD-AAU-323
47241

31 October 1985

Management Entity Annual Report
October 1, 1985 - September 30, 1986

Nutrition CRSP - Grant (USAID DAN-1309-G-SS-1070-00)

Chronology

(1 April 1986 - 31 September 1986)

April 17, 1986: Meeting of Egypt project personnel in St. Louis, MO, attended by Dr. G. Beaton and Dr. S. Murphy.

May 21-24: Visit by Dr. G. Beaton to UCLA to confer with data management and PI on Kenya data, and to Berkeley to confer with ME (Dr. D. Calloway, Dr. H. Horan, Dr. S. Murphy, Dr. J. Balderston, Dr. S. Selvin, and programmers C. Waters and D. Lein) on data management and analysis.

June 18, 1986: Visit to Berkeley by Kenya project Principal Investigator Dr. C. Neumann and researcher Ms. S. Weinberg, to confer with Dr. D. Calloway, Dr. S. Murphy, and K. Mulligan.

June 25-26, 1986: Statisticians' meeting at Purdue University, Lafayette, Indiana, including Dr. G. Beaton and Dr. S. Selvin.

July 7, 1986: Visit of Dr. S. Murphy to the University of Connecticut to confer with Dr. G. Pelto, J. Backstrand, and T. Branden on the Mexico project.

July 25, 1986: Visit of Dr. G. Beaton to Berkeley to consult with Data Management team (Dr. S. Murphy, Dr. S. Selvin, C. Waters, and D. Lein.)

July 25, 1986: Congressional presentation on all CRSPs sponsored by the National Association of State Universities and Land Grant Colleges (NASULGC), attended by Dr. G. Pelto.

August 3-5, 1986: Visit of Dr. G. Beaton to the University of Connecticut to confer with PIs, data managers, and statistician on the Mexico project.

August 7, 1986: Visit of Dr. S. Murphy and ME Data Management Research Assistant K. Mulligan to UCLA to consult with Kenya project staff.

September 12, 1986: Chicago meeting of statisticians and data managers, Principal Investigator Dr. G. Harrison, and consultants Dr. S. Fienberg and Dr. G. Beaton.

Chronology

(1 Oct. 1985 - 31 March 1986)

October 11-14, 1985: Data Analysis Group (DAG) meeting at University of Connecticut, Storrs, including project statisticians, ME consultants Dr. S. Selvin and Dr. G. Beaton, data managers, and some PIs.

November 20, 1985: Visit of Dr. S. Murphy, ME Data Manager, and C. Waters, ME programmer, to UCLA to confer with Kenya project on data management.

November 21, 1985: Brief visit of Dr. M. Forman (ST/N) to UC Berkeley.

November 26-27, 1985: Visit of Dr. S. Murphy to attend University of Kansas data management meetings.

December 16-18, 1985: Visit of Dr. G. Beaton and Dr. G. McCabe to Berkeley to consult with ME on data management procedures.

February 24-28, 1986: Meeting of the Scientific Coordination Board (SCB) at UCLA, concurrent with meetings of the statisticians, data managers, ME (Dr. D. Calloway, Dr. H. Horan, Dr. G. Beaton, Dr. S. Murphy, Dr. S. Selvin, and Business Manager K. Condon), and the Institutional Council.

March 10-14: Program effectiveness audit by AID.

March 24, 1986: Visit of Dr. S. Murphy to UCLA data management team.

(All of the above reported on in the semi-annual report.)

Table of Appended Documents

1. Dr. L. Allen's minutes of SCB Meeting 2/24-28/86.

2. Dr. D. Calloway's 4/28/86 memo on possible descriptive statistics approaches with which to describe food consumption, with RMR/BMR suggestions appended.

3. Dr. G. Beaton's minutes of meetings with UCLA staff 5/21-22/86, which focused primarily on RMR measures.

4. Dr. E. Stanek's minutes of Data Analysis Group meeting at Purdue University 6/25-26/86.

5. Documents relating to Data Analysis Group meeting in Chicago 9/12/86:

   6.1 Dr. G. Beaton's meeting notes;
   6.2 Dr. G. Harrison's trip report.

Appendices Previously Attached to Semi-Annual Report (10/1/85 - 3/31/86)

1. Policy Consultative Group:

   Dr. M. Forman's memo on PCG's proposed activities/schedule, Dr. H. Horan's response;
   Dr. Forman's memo and minutes of 2/86 SCB discussion at UCLA concerning the PCG.

2. UCLA Fiscal Problem and Resolution:

   Dr. C. Neumann's budget proposals for the period 11/85-4/86;
   Mr. P. Costic's letter and Dr. Horan's transmittal concerning budgets;
   Dr. Horan's UCLA budget memo;
   Mr. R. Edwards' IC minutes of 2/86 UCLA meeting.

3. Data Analysis Group Meeting at University of Connecticut, 10/85.

   Dr. G. Beaton's agendas;
   Dr. G. McCabe's statisticians' report;
   Dr. S. Murphy's post-meeting letters to Data Managers and related tables.

4. UCLA meetings between SCB, IC, Statisticians, and Data Managers, 2/86.

   Dr. Beaton's memo on approaches to consumer unit reduction;
   Dr. D. Calloway's list of four essential tasks in archiving and analysis;
   Dr. Horan's post-UCLA meeting general memo.
   Dr. G. McCabe's statisticians' report.

MANAGEMENT ENTITY ANNUAL REPORT SUMMARY
October 1, 1985 - September 31, 1986

The Program Coordinator's Office

The Program Coordinator's Office has strengthened its established role as organizer and integrator for the CRSP over the past year. Aside from carrying on its numerous regular functions and responsibilities--which include acting as a liaison between the various CRSP and CRSP-related groups, managing the contractual and budgetary aspects of the program, and planning various meetings and agendas for CRSP components--the office prepared visual and written materials for a Congressional presentation on all CRSPs sponsored by the National Association for State Universities and Land Grant Colleges (NASULGC), and conducted a major reorganization and cataloguing of all the books, reports, and files here at ME.

Correspondence/Communication

One of the ME office's principal functions is to facilitate communications among the various CRSP groups--Principal Investigators, statisticians, data managers, ME and its consultants, the External Evaluation Panel (EEP), the project psychologists, and the Institutional Council (IC). ME also serves as the CRSP liaison between U.S. government organizations to which the CRSP has direct ties, such as the Agency for International Development (AID) and the Board for International Food and Agricultural Development (BIFAD),as well as various voluntary and other common interest agencies and groups.

Most of the correspondence that passes through the ME office to these ends is by letter or telephone, with some affiliates making frequent use of the BITNET computer network for immediate electronic transmission of written information and requests.

During the past year, some 900 pieces of mail have been generated from within the Program Coordinator's office. Included in this batch of outgoing mail have been:

- o complex monthly mailings to PIs, statisticians, data managers, and ME consultants apprising them of current issues and concerns;
- o quarterly travel reports--including calculations for each project of cumulative total outlays, recent period outlays, and Thomas Cook Travel subcontract funds remaining;
- o ME and country project semi-annual and annual reports;
- o descriptive materials on the CRSP, usually in response to requests from interested parties.

Materials received, copied, filed, and catalogued by ME have included: numerous scientific, statistical, and data management articles and reports relevant to CRSP research; trip reports; requests from Principal Investigators for approval to release data collected for the CRSP in publications or presentations; meeting minutes, data runs, tapes, and other scientific or statistical updates.

Aside from written communication, much of ME's organization and negotiation role is fulfilled by telephone. Taking May 1986 as a random example of a typical month, the office made 170 outgoing long distance calls, and received approximately 120. Many of these calls were made in deliberating over financial matters or in planning CRSP meetings of varying sizes and purposes.

The year's major meetings included an October '85 Data Analysis Group meeting in Storrs, Connecticut; the February '86 Scientific Coordination Board (SCB) meeting in Los Angeles, involving PIs, ME, statisticians, data managers, the IC, and several participants from the field sites; a June '86 Data Analysis Group meeting at Purdue University; a September '86 Data Analysis Group meeting in Chicago; and various ad hoc smaller meetings (usually involving one or two PIs, data managers, statisticians and ME consultants or staff) at UCLA, Berkeley, and the University of Connecticut. Many of the logistical details involved in these meetings were handled by this office, even if the meeting itself was held elsewhere.

Travel

The provision of travel arrangements and reimbursements has been another significantly time-consuming component of ME's work. Because of our subcontract with Thomas Cook Travel, processing of all air travel comes through this office at one stage or another--over the past year sixty trips involving air travel were processed here.

Whether these trips were made by researchers returning from the field, or by participants in the 9 or 10 CRSP meetings held throughout the year, almost all of them have required some amount of ME follow-up. A substantial amount of time has been devoted to tracking the itineraries and currency conversions of field researchers' travel--both for home leave, and for return from the field--over the past few years. ME had withheld payment to Thomas Cook on a number of travellers' air tickets because of their deviations from AID standards (such as flying non-US carriers, or making extra long stop-overs or excursions from the standard route), for which we were awaiting clarification, justification, or reimbursement.

Occasionally, a single such airfare problem has demanded up to ten hours' working time before being resolved to the satisfaction of all parties concerned: the traveller, the funding agency, the travel agency, and the University of California travel bureau. After much work in reconstructing flight itineraries, figuring ticket costs on various currency conversion rates, review of the AID "boiler plate" on travel regulations, and discussion with various travel agency representatives, we have cleared the major part of the outstanding debts.

Financial Responsibilities

A very significant portion of this office's time and effort has been devoted, as in previous years, to overseeing and negotiating the financial and contractual arrangements of the CRSP grant.

On a regular basis, this has meant generating and sending quarterly travel reports to the country projects; negotiating periodic grant subcontract amendment requests; reviewing and balancing quarterly expenditure reports and budget summaries and projections sent in by the projects; conducting cost share and indirect cost assessment analyses; and negotiating additional funding requests.

This year we have had a number of additional funding requests from the projects for increases or extensions in staff and/or equipment. After deliberation, ME agreed to the hiring of a Research Associate to aid the Mexico project Data Manager working out of the University of Massachusetts, and to the extension of the Egypt project's Data Manager at Purdue.

Other requests for funding have been based on the need for reallocation of funds from the reserved travel budget to cover staffing needs at the

University of Arizona, and a raise in the ceiling authorization for the University of Connecticut.

The overall financial monitoring function of the Program Coordinator's office has taken on a new intensity in FY '86. As the CRSP approached its final period, budgetary decisions became at once clearer and more difficult. The project's increasingly finite time and money limitations have provoked renewed planning efforts on the part of all the projects, as well as ME.

As a first significant result of these efforts, the CRSP has prepared to enter its Period VI with significantly more funds than would have been projected a year ago: over $1.1 million still remains at UC Berkeley, and the amount of unexpended, unencumbered funds in project hands is estimated to be in excess of $200,000, for a project-wide total of $1.3 to $1.4 million.

There are two basic reasons for this economy. One, unfortunately, is that data delivery is still behind schedule in two of the three projects; costs associated with data delivery are therefore also lagging by up to three months.

More significantly, the projects have voluntarily increased cost-sharing, sometimes enormously, and have also cut costs through many intelligent fiscal choices. ME has reduced its staff considerably. The Program Coordinator's office has continued its drastic reduction in staff, having lost its Administrative Assistant in FY '85, and its Business Manager/Administrative Analyst in FY '86. This office now stands at a total of 1.80 FTE, consisting of a Coordinator and a Secretary. Additionally, the Data Management staff has been reduced by some .50 FTE.

These are all the reductions in personnel the CRSP here at UCB can manage. Indeed, more programming staff may be needed in some areas, at least for a time, as the data sets reach completion.

An additional task of some magnitude has been the handling of a series of special audits, generated essentially by the Gramm-Rudman legislation. A two-person effectiveness audit was conducted over a full week at Berkeley by an audit manager from AID. Not long after, a second two-person team visited three sites in the CRSP, including UCB, for 2-3 days per site. Their subject was cost-sharing--its advisability, and the manner and adequacy of how the established requirement is being met.

All of the audit personnel verbally pronounced themselves satisfied with the situation here, and no negative comments have been received at all.

Additionally, all of our basic financial tracking of documents and spreadsheets now exist in Lotus 1-2-3 computer files, as well as on back-up manual spreadsheets and files. As mentioned--and described in detail below--<u>all</u> files, including the financial ones, were recently reviewed, re-organized, and updated. We now believe them to be substantially ready for the final archiving process.

A major financial task during the past year, the successful resolution of a serious fiscal problem at one of the projects, is described in the semi-annual report.

Special Projects

Outside of its regularly established functions, the ME office undertook two other projects this year. One was the complete reorganization and cataloguing of all the files, books, reports, and reference materials held in the CRSP offices, both current and historical. This project developed into a major undertaking spanning a period of many months, as it involved sifting through some 900 files, 140 books and reference materials, 100 articles and expertise reports, and a dozen regularly received periodicals, all of which have been collected here at the ME offices over the past five years.

This indexing project was undertaken because we knew that the investigators, statisticians, and data managers will need to be aware of--and able to obtain--this variety of supplemental materials and support documents. ME seemed the logical place to locate this small archive of CRSP materials, both for our investigators and for AID.

The other project our office executed was the development of informational materials for a Congressional presentation on all CRSPs that was organized by NASULGC in order to promote a broader awareness of the function and scope of this unusual group of government-funded programs. This task involved designing and writing a descriptive two-page flier on the CRSP for use as a handout; selecting and developing appropriate photographs for a poster-board display from our collection of slides from the three countries; and arranging for one of our principal investigators to fly to Washington to participate in the presentation, which, according to several sources, was both successful and well-received.

Office Equipment and Personnel

As mentioned, during the past year the ME office has undergone considerable staffing changes. Dr. Judith Balderston, Associate Research Economist and coordinator for data analysis, left the CRSP to pursue other scientific and professional interests in May 1986. Also in that month, Business Manager Katherine Condon transferred to a permanent position elsewhere on the Berkeley campus. Katherine's leaving meant that all financial responsibilities had to be shifted to the Program Coordinator and the recently hired secretary. During this process, many of the record-keeping, accounting, and reporting procedures were substantially updated and redesigned, despite substantial reduction in office staff.

The ongoing upgrading of the office's microcomputer software library has helped ease this transition considerably. The acquisition of and training on the Lotus 1-2-3 software program, and a number of related programs

which expand its powers, has allowed us to improve both the speed and convenience of making budgetary and financial calculations on a broad scale, as well as the overall clarity and legibility of the records.

The data management component of the office has made good use of Lotus as well, in entering a complex nutrient database and in calculating energy values from other nutrient components of food items. This software has also been valuable for its tabulating and graphing capabilities, which have enabled us to produce clear graphs and charts for reports such as this one.

We have added several other software programs to our library as well. As adjuncts to Lotus, we acquired Note-It, Lotus 101 Macros, and Spreadsheet Auditor, all of which expand the powers and scope of the Lotus framework. A final addition has been that of a Hayes Smartcom Modem, which has enabled the ME office to communicate with the UC Berkeley mainframe and thereby send BITNET messages across the country. Most of our staff have also achieved competency in Wordperfect, now the Nutrition CRSP's standard word-processing package.

ME DATA MANAGEMENT AND ANALYSIS REPORT

Data management and analysis activities have focused primarily on archiving this past year. The majority of the field data from all three projects has been received in Berkeley and copied, and the status of the files tabulated. In addition, we have begun documenting the data sets and reorganizing them into a final archive format. These activities have required extensive consultation with the staffs of the three projects, often via BITNET (electronic mail), frequently via memos and letters, and occasionally via travel to the project locations.

Our current staff consists of a data manager/nutritionist (Dr. Suzanne Murphy), two programmers (David Lein and Claudia Waters), a statistician (Prof. Steve Selvin), and a research assistant (Kathleen Mulligan), working a total of 2.3 FTE, of which .7 FTE is contributed by the UCB Nutritional Sciences Department.

Status of Archived Files

Tables 1 through 3 show the files that have been received for archiving from each of the three projects, and the number of records on each file. The left-most column indicates the names of the 32 files that were defined by the September 1984 Basic Data Set. Thus, it is possible to determine how the archived project files relate to the originally defined files.

The next step will be to expand these tables, using mapping forms supplied by each project, and to relate the variables on each file to the originally defined Basic Variables. When this task is completed it will be possible for researchers to determine which Basic Variables are available across multiple projects.

The two right-hand columns of Tables 1 through 3 give more detail on the completeness of each project's data sets. For each file archived, the dates (beginning and ending) of the observations on the file are reported. The assumed status of each file is shown in the last column. Many of the files are now complete, others are awaiting final checking by the US institutions, and a few are still in the host countries pending data entry, checking, or organization. Files with derived variables (such as nutrient data, SES scores, indices of morbidity, etc.) are still partially incomplete for all projects. In several cases, these calculations were postponed until the field data sets were complete.

Format of the Final Archive

The ME data management staff has begun working on the format of the final archive. The archived files must encompass two aspects of the final data sets: the similarities between the original design and the actual data of the three field projects, and the dissimilarities imposed by the realities of the actual field work.

Users of the CRSP archived data will need to clearly understand the cross-project similarities and differences in each variable, and, in some cases, changes in the collection methodology within a variable over time. Thus, it is critical that the archived files be clearly documented, that similarities be indicated when appropriate, and that differences be described as well.

We propose archiving three levels of file structure:

1. The field data sets, reflecting the data as originally entered onto tapes (usually in the host country). These data would have been range-checked and corrected. The files would not necessarily be organized into a logical structure, but they might reflect the chronological nature of the collection and entry process. Future researchers might wish to use these files to investigate the actual values of variables prior to correction or transformation by the U.S. institutions. These data would also be useful for examining changes over time in data entry procedures, etc. Documentation would be provided by the projects.

2. Data sets archived by the U.S. institutions, with necessary changes made by the ME staff for structuring files into the Basic Variable format. These data sets would be in SAS format, with similar file and variable names as appropriate. Project-specific files would not be transformed by ME, but would be archived as received. Documentation would be provided by the projects and updated as necessary by ME.

3. Analytic data sets, created either by the projects or ME, for the purpose of performing a specific analysis. These data sets could be used by future researchers to replicate analyses presented in the various CRSP reports. Documentation would be provided by the creator of the data set.

Of the three levels, only level two would be required for the official archive. However, ME will collect and archive files at levels one and three, as available.

Descriptive Statistics of Basic Variables

An important part of ME's archiving task is the description and documentation of the contents of each data set. To this end, we have begun with descriptive statistics of some of the more important basic variables. For these initial reports, we chose to work with three files: food intake, anthropometry, and morbidity. We also chose a subset of the variables on

each file that would be relatively comparable across all three projects. Thus, we created the following three SAS files for each project (for a total of nine SAS data sets):

Food intake file: variables = energy intake, date of measure.

Anthropometry file: variables = height, weight, date of measure.

Morbidity file: variables = begin and end dates of illness, illness code, severity code.

In addition, each observation contains a standard "header": household ID, target type, and birth date.

We then used SAS to obtain descriptive statistics on these variables (mean, median, standard deviation, and ranges for continuous variables; frequencies on categorical variables). Some of these results are shown in Tables 4-6 (anthropometry) and 7 (morbidity). The energy intakes are plotted by calendar month and target type for each project in Figures 1 through 3. The data for these plots are shown in Tables 8-10.

Although these tables may appear to have been rather simple to generate, each required substantial thought and definitions. For example, a first step was to define the time window that was acceptable for each target type to be described. For the anthropometry tables, it was necessary to ensure that the infants and toddlers were approximately the same age in all three projects, so that heights and weights could be compared. A window was established for each age group (see footnotes), and only children who had measurements taken in this window were described by the statistics. In some cases, the number of targets appears artificially low due to the incompatibility of the collection methodology and the definition of the time window.

Each project defined morbidity codes that would reflect those illnesses that are prevalent in the host country. Thus, comparing illness types across countries presented some difficulties. The illness categories shown in Table 7 required summarization of similar illness codes among projects, as well as an analysis of those illnesses that were prevalent in each country. (See footnotes for details.)

Projected Data Available For Analyses

Note that most of these data are cross-sectional in nature. We have made no attempt to ensure that the individuals on these files are truly targets as defined by the CRSP protocol--that is, those belonging to a cluster containing an index individual plus associated household members, as shown below:

Index Individual	Associated Individuals
Pregnant woman entering second trimester	Lead male, infant 0-6 months
Toddler at 18 months	Lead male, lead female
Schooler, 7-9 years	Lead male, lead female

Future longitudinal descriptions of data should focus on only those individuals in one of these three clusters.

Furthermore, to be a target in longitudinal analyses, an individual should have been observed a minimum number of times across the year of observation. Table 14 proposes a method of describing the completeness of a data file, using the Kenya Project food intake file for toddlers. Researchers will have to define the minimum number of observations, as well as the maximum number of skipped months, that they require for their analyses.

TABLE 1
CROSSREFERENCE, MEXICO FILES TO BASIC FILES; STATUS OF ARCHIVED FILES

BASIC FILE	MEXICO NAME	NUMBER ARCHIVED	DATES OF ARCHIVED DATA	ASSUMED FILE STATUS
Entry/exit/change	Basal census	5631	N/A	Being Updated
	Sample census	1461	N/A	Being Updated
Mortality	see Clinic. Hist.			all data in Mexico
Household intake	see Diet (raw)			
Individual intake	Diet (kcals)	12879	1/84-5/85	6/85-2/86 at UConn
	Diet (raw)	370897	1/84-2/86	3/86-5/86 in Mexico Sched. for 10/86
Supplemental feeding	see Diet			
Anthropom-target	Anthropom	6689	1/84-5/86	Complete
Anthropom-non-target	HH Anthropometry	669	2/86-5/86	Complete
Metabolic adaptation	RMR	1874	1/84-5/86	Complete
Disabilities/ chronic illness	Clinical history	none		Being Entered
Reproductive history	see Clinical hist			
Physical assessment	Physical exam	none		Being Entered
Pregnancy outcome	see Physical exam			
Conception survey	see Clinic. Hist.			
Pregnancy survey	see Clinic. Hist.			
Lactation/ infant feeding	see Clinical hist			
Lab assessment/ immunology	Urine	9085	2/84-3/86	Complete
	Feces	2618	2/84-3/86	Complete?
Morbidity episodes	Morbidity	36570	1/84-5/86	Complete
Morbidity summary	see Morbidity			
Adult cognitive	Adult-cognitive	412	1/84-1/86	Complete?
Infant cognitive	Psych-6 mo.	75	1/84-1/86	"
Toddler cognitive	Psych-18 mo.	101	1/84-1/86	"
	Psych-24 mo.	118	1/84-10/85	"
	Psych-30 mo.	69	1/84-1/86	"
Schooler cognitive	Cognitive-school	291	1/84-1/86	"

TABLE 1
CROSSREFERENCE, MEXICO FILES TO BASIC FILES; STATUS OF ARCHIVED FILES

BASIC FILE	MEXICO NAME	NUMBER ARCHIVED	DATES OF ARCHIVED DATA	ASSUMED FILE STATUS
Infant behavior	Psych-3 mo.	87	1/84-1/86	"
	Brazelton (also see Psych-6 mo)	61	1/84-11/85	"
Toddler behavior	see Psych-18 mo, Psych-24 mo, Psych-30 mo			
Schooler behavior	Classroom	211	1/84-1/86	"
	Playground	270	1/84-1/86	"
HH sanitation/ hygiene	Sanitation/hygien	8920	1/85-2/86	Complete?
Individual SES	see Socioeconomic and Sociocult.			
Demography	Migration (see also Sample Census, Basal census)	439	1/84-5/85	6/85-5/86 in Mexico
Household SES	Socioeconomic	439	1/84-6/85	7/85-5/86 in Mexico
	Sociocultural	558	1/84-5/85	6/85-5/86 in Mexico
Community climate	Climatic Data	none		Being Entered
Childcare/ sanitation, LF	see Sanitation/hygiene			
Time allocation	Activity: LM	6675	1/84-12/84	Complete
	Activity: LF	56276	1/84-12/84	Complete
	Activity: family	111	1/84-12/84	Complete
Project specific files:	Weekly food use	54299	1/84-2/86	Complete?
	Food prices	none		Being Entered
	Diet substudy	none		Being Entered
	Morbidity Substudy	none		Being Entered
	Household Productivity	none		Sched. for 10/86

TABLE 2
CROSSREFERENCE, EGYPT FILES TO BASIC FILES; STATUS OF ARCHIVED FILES

BASIC FILE	EGYPT NAME	NUMBER ARCHIVED	DATES OF ARCHIVED DATA	ASSUMED FILE STATUS
Entry/exit/change	Entry/exit/change	223	N/A	8/84-12/85 in Kansas
Mortality	Mortality	none		Forms in AZ
Household intake	Household intake (unadjusted)	4209	12/83-5/85	6/85-12/85 in Kansas
Individual intake	Individual intake (kcal)	16077	12/83-12/85	Complete
Supplemental feeding	Supplemental feeding-infant	none		9/84-10/85 in Egypt
Anthropometry - target	Anthropometry - target	8030	11/83-12/85	Complete
	Anthropometry - pregnant	766	11/83-8/85	6/85-12/85 in KS
	Anthropometry - lactating	746	7/84-11/85	6/85-12/85 in KS
	Anthropometry - infant	942	1/84-12/85	6/85-12/85 in KS
Anthropom - non-target	Anthropom - non-target	603	11/83-5/85	Complete
Metabolic adaptation	RMR	798	10/84-10/85	11/85-12/85 in Kansas
Disabilities/ chronic illness	Medical history	1274	12/83-8/84	8/84-12/85 in KS
Reproductive history	Reproductive history	118	1/84-7/84	8/84-12/84 in KS 6/85-1/86 in Egypt
	Reproductive history lactation history	none		12/84-1/86 in KS
Physical assessment	Physical exam	1219	12/83-9/84	8/84-12/85 in KS
Pregnancy outcome	Pregnancy outcome	124	4/84-5/85	Complete
Conception survey	see Reproductive history			
Pregnancy survey	Pregnancy monthly visit	50		8/84-3/85 in KS
Lactation/ infant feeding	Lactation/infant feeding practices	3	7/84-8/84	8/84-3/85 in KS 4/85-9/85 in Egypt
Lab assessment/ immunology	Hematology/urine	3931	11/83-12/85	Complete
	Parasitology	656	7/84-8/84	1&2/84, 1&2/85 in KS
	Immunology	655	11/83-12/84	4/84-6/85 in KS
	Biological	532	11/83-7/84	Complete
Morbidity episodes	Morbidity illness episodes	3888	10/83-12/85	Complete
Morbidity summary	Morbidity weekly recall	88921	10/83-12/85	Complete

TABLE 2
CROSSREFERENCE, EGYPT FILES TO BASIC FILES; STATUS OF ARCHIVED FILES

BASIC FILE	EGYPT NAME	NUMBER ARCHIVED	DATES OF ARCHIVED DATA	ASSUMED FILE STATUS
Adult cognitive	Father cognitive	20	12/83-5/84	8/84-12/85 in Egypt
	Mother cognitive	44	12/83-5/84	8/84-12/85 in Egypt
Infant cognitive	IBR scores	none		Data in Egypt
Toddler cognitive	Toddler cognitive	53	12/83-5/84	8/84-12/85 in Egypt
Schooler cognitive	Schooler cognitive	41	12/83-5/84	8/84-12/85 in Egypt
Infant behavior	Infant behavioral	246	3/85-10/85	Complete
	Dubowitz	110	7/84-5/85	Complete
	Brazelton	none		7/84-9/85 inEgypt
Toddler behavior	Toddler behavior	2586	12/83-11/85	Complete
Schooler behavior	Schooler classroom	113	12/83-4/84	8/84-12/85 in Egypt
	Child behavior ratings	25	2/84-3/84	2/85-4/85 in Egypt
HH sanitation/ hygiene	HH sanitation/hygiene (scores)	188	1/84-9/84	8/84-10/84 in KS 11/84-12/85 in Egypt
	Toddler HH sani/hygiene (scores)	156	1/84-9/84	8/84-10/84 in KS 11/84-12/85 in Egypt
	Toddler indiv sani/hygiene (scores)	150	1/84-9/84	8/84-10/84 in KS 11/84-12/85 in Egypt
Individual SES	Individual SES	2180	N/A	Complete
Demography	Demography	2180	N/A	Complete
	Demographic update	none		8/84-9/84 in KS Remaining data in Egypt
Household SES	Household SES - scores	312	N/A	Complete
	Household SES	191	10/83-10/84	Complete
	Household SES - Update	none		11/84-12/85 in KS
Community climate	Community Climate	none		11/83-12/85 in Egypt
Childcare/ sanitation, LF	Childcare/sanitation, LF	none		1/84-8/84 in KS Remaining data in Egypt
Time allocation	Time allocation	none		7/84-10/84 in KS Remaining data in Egypt
Project specific files	Water microbiology	none		8/84-7/85 in KS 8/85-12/85 in Egypt
	Breastfed toddlers	2799	12/83-11/85	Complete
	Number of meals consumed by lead male	4763	12/83-12/85	Complete
	Anthropometry substudy	none		6/85-12/85 in KS

TABLE 2
CROSSREFERENCE, EGYPT FILES TO BASIC FILES; STATUS OF ARCHIVED FILES

BASIC FILE	EGYPT NAME	NUMBER ARCHIVED	DATES OF ARCHIVED DATA	ASSUMED FILE STATUS
	Physiologic state of target female	5150		
	HH food and beverage frequency	none		
	Food and beverage prices (SESCO)	9	12/83-5/85	3/84,6/85-9/85, 11/85-12/85 in KS 11/83, 4/84-11/84, 10/85 in Egypt
	Wages	none		4/85, 7/85 in KS 11/83-3/85, 5/85-6/85, 8/85-12/85 in Egypt
	Prices of locally grown food	none		4/85-8/85 in KS 11/83-3/85, 9/85-12/85 in Egypt

TABLE 3
CROSSREFERENCE, KENYA FILES TO BASIC FILES; STATUS OF ARCHIVED FILES

BASIC FILE	KENYA NAME	NUMBER ARCHIVED	DATES OF* ARCHIVED RECORDS	ASSUMED FILE STATUS
Entry/exit/change	Entry/exit/change	1130	N/A	Complete
Mortality	Mortality notification	4	N/A	"
Household intake	Meal prep. summary	38845	All	"
	Consumer unit	37502	All	"
	Recipe nutrients	38845	All	"
Individual intake	TI/NTI summary	26675	All	"
	Indiv nutr/hand calc	6014	3/84-5/85	"
	Indiv nutrients	24707	All	"
Supplemental feeding	See lactation quest.			
Anthropom-target	Anthropom	14268	All	"
non-target	see Anthropom			
Metabolic adaptation	RMR	4712	All	"
Disabilities/ chronic illness	Disability/ chronic disease	727	1/84-11/85	"
Reproductive history	Reproductive hist	288	1-2/84	"
Physical assessment	Clinical exam-summary	2162	All	"
Pregnancy outcome	Pregnancy outcome	138	3/84-7/85	"
Conception survey	see Pregnancy survey			
Pregnancy survey	Pregnancy survey	5857	1/84-9/85	"
Lactation/ infant feeding	Lactation question.	879	5/84-12/85	"
Lab assessment/ immunology	Lab-physiol. samples	2136	2/84-12/85	"
	Immunology/water contamin.	1025	4/85-8/85	"
Morbidity episodes	Morbidity-indiv 4 week	5788	2/84-11/84	"
	Morbidity-indiv 4 week	8764	All	"
	Morbidity-HH 4 week	8970	All	"
Morbidity summary	not computerized			
Adult cognitive	Cog adult summary	674	7/84-12/85	"
Infant cognitive	Cog infant summary	312	All	"
Toddler cognitive	Cog toddler summary	315	1/84-4/85	"
	Cog toddler-30 mo.	230	12/84-11/85	"
Schooler cognitive	Cog schooler summary	462	6/84-8/85	"

TABLE 3
CROSSREFERENCE, KENYA FILES TO BASIC FILES; STATUS OF ARCHIVED FILES

BASIC FILE	KENYA NAME	NUMBER ARCHIVED	DATES OF* ARCHIVED RECORDS	ASSUMED FILE STATUS
Infant behavior	Infant interaction	6406	All	"
	Infant behavior	131	2/84-7/85	"
Toddler behavior	Toddler interaction	9933	1/84-10/85	"
Schooler behavior	Schooler classroom observation	494	5/84-9/85	"
	Schooler playground observation	1516	4/84-11/85	"
	School attendance/ performance	211	12/84,12/85	"
HH sanitation/ hygiene	Sanitation/hygiene	1305	1/84-10/85	"
	Sanitation/hygiene	424	7/85-10/85	"
Individual SES	see Census update and SES			
Demography	Census update	11505	1/84-11/85	"
Household SES	SES	1402	All	"
Community climate	Weather data	566	7/84-12/85	"
Childcare/ sanitation, LF	Care-giving activities	32471	All	"
Time allocation	Time allocation	13978	N/A	"
Project specific:	Adult Literacy test	524	5/85-12/85	"
	Agricultural crop questionnaire	6963	All	"
	Market survey summary	131	5/84-2/85	"
	Depression/alcohol questionnaire	499	9-12/85	"

First and last dates;
All : 1/84 - 12/85

TABLE 4: CROSS PROJECT COMPARISON OF ANTHROPOMETRIC MEASURES FOR SELECTED TARGETS: WEIGHT (KGs)

	EGYPT WEIGHT				KENYA WEIGHT				MEXICO WEIGHT			
	MED	MEAN	STD	N	MED	MEAN	STD	N	MED	MEAN	STD	N
Infants-Birth	3.5	3.4	0.05	*113								
Male					3.2	3.3	0.5	71	3.2	3.2	0.4	49
Female					3.1	3.2	0.4	57	3.0	3.1	0.4	45
Infants-6 mos	6.6	6.7	1.0	*71								
Male					7.0	7.1	0.9	63	7.2	7.2	1.2	50
Female					6.7	6.8	1.1	51	6.5	6.6	0.9	45
Toddlers-18 mos												
Male	10.2	11.0	6.1	55	9.4	9.5	7.1	42	9.3	9.5	0.9	26
Female	9.9	9.5	1.5	40	8.7	7.4	3.7	49	9.7	9.7	1.3	23
Toddlers-24 mos												
Male	11.6	11.5	1.8	31	10.6	10.5	1.1	52	10.4	10.6	1.0	56
Female	10.6	10.9	1.7	34	9.9	10.0	1.1	55	10.4	10.6	1.3	55
Toddlers-30 mos												
Male	12.8	12.9	1.8	38	11.7	11.6	1.2	43	11.6	11.6	1.1	54
Female	12.2	13.0	6.5	51	11.0	10.9	1.2	46	11.6	11.7	1.4	43
Schoolers-Entry	22.6	23.0	5.5	*133								
Male					19.0	19.2	2.8	91	21.3	21.8	3.5	92
Female					19.4	19.3	2.8	75	20.1	20.5	2.8	84
Adults-Entry												
Male	66.1	71.6	37.8	191	54.6	55.6	7.2	242	64.5	66.2	9.8	213
Female	62.5	65.0	11.8	209	50.1	51.4	8.7	283	57.7	58.3	9.3	275

* Sex combined

TABLE 5: CROSS-PROJECT COMPARISON OF ANTHROPOMETRIC MEASURES FOR SELECTED TARGETS: HEIGHT (CMS)

	EGYPT HEIGHT				KENYA HEIGHT				MEXICO HEIGHT			
	MED	MEAN	STD	N	MED	MEAN	STD	N	MED	MEAN	STD	N
Infants-Birth	51.0	51.0	2.0	*42								
Male					49.3	49.7	2.5	46	50.0	50.1	2.2	39
Female					48.5	48.6	2.5	40	49.0	48.9	2.1	42
Infants-6 mos	64.9	77.5	89.7	*36								
Male					62.6	62.6	2.7	63	64.3	64.4	2.6	50
Female					61.9	61.7	2.7	51	62.3	62.2	2.0	45
Toddlers-18 mos												
Male	76.7	76.5	3.8	50	74.9	75.8	3.6	42	75.1	74.7	4.0	18
Female	75.0	74.3	5.3	39	74.1	74.0	5.9	50	76.1	75.6	3.7	14
Toddlers-24 mos												
Male	80.4	80.9	3.9	30	80.1	79.8	3.6	52	79.1	80.0	3.7	22
Female	79.5	78.8	3.3	33	78.3	78.7	2.9	55	78.6	80.0	4.1	12
Toddlers-30 mos												
Male	85.3	84.1	3.3	35	83.6	83.8	3.6	43	82.2	83.1	3.4	20
Female	83.0	82.8	4.4	49	82.1	82.8	3.0	46	85.1	85.2	3.1	10
Schoolers-Entry	119.2	119.7	8.0	*134								
Male					113.2	113.3	5.9	89	118.3	117.8	6.7	88
Female					113.0	113.6	6.4	74	115.6	115.8	6.1	79
Adults-Entry												
Male	167.8	166.1	11.3	111	165.6	165.4	6.5	239	166.0	166.0	5.8	213
Female	155.1	155.2	8.1	171	154.2	154.5	5.7	278	152.5	152.8	5.4	213

* Sex combined

TABLE 6: CROSS-PROJECT COMPARISON OF ANTHROPOMETRIC MEASURES FOR SELECTED TARGETS: WEIGHT/HEIGHT

	EGYPT (KGs/CMs)				KENYA (KGs/CMs)				MEXICO (KGs/CMs)			
	MED	MEAN	STD	N	MED	MEAN	STD	N	MED	MEAN	STD	N
Infants-Birth	0.07	0.08	0.01	*35								
Male					0.07	0.07	0.01	46	0.06	0.06	0.01	39
Female					0.07	0.07	0.01	40	0.06	0.06	0.01	42
Infants-6 mos	0.11	0.11	0.03	*27								
Male					0.11	0.11	0.01	59	0.11	0.11	0.01	50
Female					0.11	0.11	0.02	48	0.10	0.11	0.01	45
Toddlers-18 mos												
Male	0.14	0.13	0.01	49	0.12	0.13	0.10	42	0.12	0.13	0.01	18
Female	0.13	0.13	0.02	39	0.12	0.10	0.05	49	0.12	0.13	0.02	14
Toddlers-24 mos												
Male	0.14	0.14	0.02	29	0.13	0.13	0.01	52	0.13	0.13	0.01	22
Female	0.14	0.14	0.02	32	0.13	0.13	0.01	55	0.12	0.13	0.01	12
Toddlers-30 mos												
Male	0.15	0.15	0.01	34	0.14	0.14	0.01	43	0.14	0.13	0.01	20
Female	0.15	0.16	0.08	49	0.13	0.13	0.01	46	0.14	0.14	0.01	10
Schoolers-Entry	0.19	0.19	0.03	*130								
Male					0.17	0.17	0.02	91	0.18	0.18	0.02	88
Female					0.17	0.17	0.02	75	0.18	0.18	0.02	79
Adults-Entry												
Male	0.40	0.04	0.30	110	0.33	0.34	0.04	238	0.39	0.40	0.05	213
Female	0.40	0.42	0.08	168	0.32	0.33	0.05	278	0.38	0.38	0.06	213

* Sex combined

ANTHROPOMETRY FOOTNOTES (TABLES 4-6)

Egypt - Table 4:

1. Statistics for toddlers, schoolers, and adults based upon measurements taken between February 1984 and August 1985; statistics for infants based upon measurements taken between January 1984 and December 1985.

2. 'Infants at Birth' refers to the first measurement for each infant between the interval of -.5 and .9 months.

3. 'Infants at 6 Mos' refers to the first measurement for each infant between the interval of 5.5 and 6.9 months.

4. 'Toddlers at 18 Mos' refers to the first measurement of each toddler between the interval of 17.5 and 18.9 months.

5. 'Toddlers at 24 Mos' refers to the first measurement of each toddler between the interval of 23.5 and 24.9 months.

6. 'Toddlers at 30 Mos' refers to the first measurement of each toddler between the interval of 29.5 and 30.9 months.

7. 'Schoolers at Entry' refers to the first measurement of each schooler.

8. 'Adults at Entry' refers to the first measurement of each adult.

9. 'Adult Females' includes pregnant and lactating females.

Kenya - Table 5:

1. Statistics based upon measurements taken between January 1984 and December 1985.

2. 'Infants at Birth' refers to the first measurement for each infant between the interval of -.5 and .9 months.

3. 'Infants at 6 Mos refers to the first measurement for each infant between the interval of 5.5 and 6.9 months.

4. 'Toddlers at 18 Mos' refers to the first measurement of each toddler between the interval of 17.5 and 18.9 months.

5. 'Toddlers at 24 Mos' refers to the first measurement of each toddler between the interval of 23.5 and 24.9 months.

6. 'Toddlers at 30 Mos' refers to the first measurement of each toddler between the interval of 29.5 and 30.9 months.

Kenya Anthropometry Footnotes (cont.) - Table 5

7. 'Schoolers at Entry' refers to the first measurement of each schooler.

8. 'Adults at Entry' refers to other first measurement of each adult.

9. 'Adult Females' may include pregnant and lactating females.

10. During the period of January 1984 to April 1985, daily records for height and weight are the mean of two measurements, taken at the same examination.

Mexico - Table 6

1. Statistics based upon measurements taken between March 1984 and May 1986.

2. 'Infants at Birth' refers to the first measurement for each infant between the interval of -.5 and .9 months.

3. 'Infants at 6 Mos' refers to the first measurement for each infant between the interval of 5.5 and 6.9 months.

4. 'Toddlers at 18 Mos' refers to the first measurement of each toddler between the interval of 17.5 and 18.9 months.

5. 'Toddlers at 24 Mos' refers to the first measurement of each toddler between the interval of 23.5 and 24.9 months.

6. 'Toddlers at 30 Mos refers to the first measurement of each toddler between the interval of 29.5 and 30.9 months.

7. 'Schoolers at Entry' refers to the first measurement of each schooler.

8. 'Adults at Entry' refers to the first measurement of each adult.

9. 'Adult Females' may include pregnant and lactating females.

TABLE 7: CROSS-PROJECT COMPARISON OF MORBIDITY MEASUREMENTS FOR TODDLERS

ILLNESS DURATION IN DAYS

	EGYPT				KENYA				MEXICO			
	MED	MEAN	STD	N	MED	MEAN	STD	N	MED	MEAN	STD	N
MALE	4	5.2	4.3	697	7	10.1	8.1	1270	6	6.1	3.6	358
FEMALE	4	5.4	4.1	664	7	10.5	8.6	1650	6	6.1	3.6	266

ILLNESS SEVERITY (%)

	EGYPT			KENYA			MEXICO		
	MILD	SEVERE	N	MILD	SEVERE	N	MILD	SEVERE	N
MALE	43.6	56.4	700	94	6	1268	97	3	379
FEMALE	40.2	59.8	665	93	7	1645	97	3	269

TYPE OF MORBIDITY--SEX COMBINED (%)

	EGYPT N=1455	KENYA N=2920	MEXICO N=647
COMMON COLD	10	27	33
ACUTE UPPER RESP	5	15	5
LOWER RESP	12	2	21
DIARRHEA	25	8	17
FEVER, MALARIA	9	6	13
CONJUNCTIVITIS	9	21	1
OTHER	29	21	11

TODDLER MORBIDITY FOOTNOTES (TABLE 7)

Egypt:

1. Statistics are based upon measurements taken between October 1983 and December 1985 for all toddlers showing an illness episode.

2. For the 'type of morbidity' measurement, 'common cold' refers to illnesses coded as 'common cold' or 'chronic purulent rhinitis'; 'acute upper respiratory' refers to illnesses coded as 'tonsillo-pharyngitis'; 'lower respiratory' refers to illnesses coded as 'croup', 'bronchitis', 'asthma', or 'pneumonia'; 'diarrhea' refers to illnesses coded as 'diarrhea'; 'fever/malaria' refers to illnesses coded as 'fever, no other symptoms', 'fever, other symptoms', 'fever, post immunization', 'malaria', or 'rheumatic fever'; 'conjunctivitis' refers to illnesses coded as 'conjunctivitis'; 'other' refers to all other illness categories (residual).

Kenya:

1. Statistics based upon measurements taken between January 1984 and December 1985 for all toddlers showing an illness episode.

2. For the 'type of morbidity' measurement, 'common cold' refers to illnesses coded as 'common cold'; 'acute upper respiratory' refers to illnesses coded as 'tonsillo-pharyngitis' or 'acute upper respiratory infection'; 'lower respiratory' refers to illnesses coded as 'group/laryngotracheitis', 'bronchitis', 'asthma/wheezing', or 'pneumonia'; 'diarrhea' refers to illnesses coded as 'diarrhea'; 'conjunctivitis' refers to illnesses coded as 'conjunctivitis'; 'other' refers to all other illness categories (residual).

Mexico:

1. Statistics based upon measurements taken between January 1984 and May 1986 for all toddlers showing an illness episode.

2. Data reflect only illnesses present on the day of the interview.

3. For the 'illness severity' measurement, 'severe' refers to states of incapacity coded as 'bedridden'.

4. For the 'type of morbidity' measurement, 'common cold' refers to illnesses coded as 'flu, cold'; 'acute upper respiratory' refers to illnesses coded as

TABLE 8
EGYPT FOOD INTAKE BY TARGET TYPE AND SEX
MEAN KCALS/MONTH

	LM	n	LF	n	MS	n	FS	n	MT	n	FT	n
12/83	2268	73	2056	76	1621	12	1763	13	804	31	707	22
1/84	2030	64	2178	65	1543	10	1744	18	685	28	738	28
2/84	2126	91	2047	94	1705	7	1591	18	842	38	931	46
3/84	2423	165	2157	171	2022	29	1875	32	876	68	897	64
4/84	2240	148	2135	159	1719	33	1717	23	971	63	896	66
5/84	2437	180	2240	199	1829	29	1745	33	1005	56	951	65
6/84	2400	214	2117	219	1924	26	1907	38	938	69	969	67
7/84	2345	240	2083	270	1810	52	1853	53	1044	91	1132	77
8/84	2289	245	2070	267	1619	49	1686	54	1069	90	1099	84
9/84	2384	249	2185	272	1903	36	1960	46	1194	83	1458	57
10/84	2319	307	2117	334	1731	57	1789	54	1139	91	1148	85
11/84	2263	321	2257	350	1865	67	1802	77	1199	112	1153	90
12/84	2404	354	2223	378	1743	81	1695	65	1214	115	1289	100
1/85	2275	315	2072	342	1689	72	1574	77	1174	103	1226	89
2/85	2240	333	2084	344	1733	69	1757	81	1080	90	1181	73
3/85	2482	310	2289	345	2025	73	1824	78	1233	86	1211	67
4/85	2414	260	2097	302	1866	61	1726	80	1140	59	1101	66
5/85	2444	217	2169	240	1790	44	1924	68	1348	46	1216	46
6/85	2545	183	2191	201	1754	35	1923	58	1259	32	1198	42
7/85	2449	122	2166	137	1746	29	1743	40	1201	21	1079	29
8/85	2712	79	2459	86	2032	20	1847	24	1386	13	1095	19
9/85	2620	66	2381	71	1994	15	2049	31	1268	13	1119	15
10/85	2529	45	2261	49	1905	17	1921	20	1189	10	1115	14
11/85	2665	26	2388	30	2113	8	1899	18	1217	6	1372	8
12/85	2993	8	2300	8			2172	8				

TABLE 9
MEXICO FOOD INTAKE BY TARGET TYPE AND SEX
MEAN KCALS/MONTH

	LM	n	LF	n	MS	n	FS	n	MT	n	FT	n
12/83												
1/84	2769	142	2499	172	1739	67	1542	60				
2/84	2744	165	2275	203	1642	74	1630	75				
3/84	2706	179	2193	225	1792	88	1492	70			514	1
4/84	2732	135	2210	151	1808	75	1508	59	688	7	780	8
5/84	2638	114	2146	148	1786	68	1656	66	515	19	851	5
6/84	2753	193	2184	273	1560	144	1289	123	792	28	1342	15
7/84	2794	184	2339	266	1596	140	1472	104	964	29	1006	19
8/84	2895	58	2296	69	1829	29	1914	27	612	5	693	14
9/84	2774	73	2046	93	1449	34	1363	36	904	18	836	16
10/84	2774	246	2434	338	1895	120	1692	111	935	74	1161	61
11/84	3048	288	2471	360	2018	116	1642	103	1057	77	1075	77
12/84	3052	273	2508	354	1847	122	1598	109	879	77	971	79
1/85	3299	212	2769	291	2183	104	2103	92	1088	64	1307	43
2/85	3337	242	2715	352	2199	110	1977	109	1128	92	1278	77
3/85	3365	246	2830	346	2143	95	1920	104	1180	92	1074	82
4/85	3058	224	2673	334	2104	96	1902	102	1055	84	1155	76
5/85	3275	267	2756	371	2301	102	1923	98	1149	92	1093	85

TABLE 10
KENYA FOOD INTAKE BY TARGET TYPE AND SEX
MEAN KCALS/MONTH

	LM	n	LF	n	MS	n	FS	n	MT	n	FT	n
12/83												
1/84	1734	130	1761	139	1596	50	1414	39	788	18	591	24
2/84	1974	366	1721	380	1521	114	1251	92	588	60	627	75
3/84	2163	349	1870	374	1540	118	1357	88	708	71	615	85
4/84	2036	385	1649	407	1604	113	1493	88	715	75	669	88
5/84	1863	433	1626	455	1457	135	1198	115	685	80	645	98
6/84	1938	425	1614	445	1483	126	1379	102	781	80	685	92
7/84	1921	440	1668	458	1510	136	1420	115	736	86	721	100
8/84	1705	449	1482	464	1446	137	1279	117	786	87	765	94
9/84	1550	435	1302	458	1215	144	1083	117	808	84	658	94
10/84	1594	427	1357	456	1224	150	1186	124	843	84	674	96
11/84	1568	435	1298	453	1264	149	1180	121	717	86	732	95
12/84	1647	418	1468	441	1270	146	1270	119	830	88	729	93
1/85	1922	408	1725	433	1622	147	1379	123	774	88	802	95
2/85	1912	331	1580	353	1541	81	1168	76	817	85	903	91
3/85	2097	315	1855	329	1596	68	1366	67	887	81	844	73
4/85	1957	271	1670	286	1497	59	1346	48	873	54	805	48
5/85	1890	266	1719	283	1489	53	1404	43	776	35	1003	32
6/85	2095	253	1736	278	1778	47	1407	42	945	24	811	20
7/85	1929	261	1770	278	1556	36	1332	35	1066	14	752	10
8/85	1885	259	1617	282	1400	26	1247	22	586	10	748	6
9/85	1992	251	1799	269	1836	28	1304	14	630	8	976	4
10/85	2133	254	1790	276	1665	16	1442	12	976	6	1055	2
11/85	1967	160	1779	181	1350	14	1089	10				
12/85	1740	14	1796	16	1640	4	1126	2				

FOOD INTAKE FOOTNOTES (TABLES 8-10, FIGURES 1-3)

LM = Lead male
LF = Lead female
MS = Male schooler
FS = Female schooler
MT = Male toddler
FT = Female toddler

Egypt - Table 8, Figure 1:

1. Statistics are based on measurements taken between December, 1983 and December, 1985.

2. Records showing daily intake exceeding 10,000 kcals have been eliminated as coding errors (this affects two records).

3. Toddler measurements include breast-fed toddlers.

4. 'Lead Females' includes pregnant and lactating females.

5. Intakes for Lead Males may be low if meals were consumed away from home.

Mexico - Table 9, Figure 2:

1. Statistics based upon measurements taken between January 1984 and May 1985. This is a preliminary file; the number of observations will increase and the intake values may change when the final file is available.

2. 'Lead Females' includes pregnant and lactating females.

3. Food intake collection methodology was changed in October, 1984 (new forms) and in December, 1984 (different personnel). The impact of these changes has been discussed in previous Mexico Project reports.

Kenya - Table 10, Figure 3:

1. Statistics are based upon measurements taken between January 1984 and December 1985.

2. 'Lead Females' includes pregnant and lactating females.

3. Toddler measurements include some breast-fed toddlers.

4. Records showing daily intake exceeding 10,000 kcals have been excluded (affects 13 records).

Mexico Toddler Morbidity Footnotes (cont.) - Table 7

flu, cold'; 'acute upper respiratory' refers to illnesses coded as 'throat infection', 'sore throat', or 'laryngitis'; 'lower respiratory' refers to illnesses coded as 'cough', 'chills', 'bronchitis', or 'hervor de pecho' (translation not available); 'diarrhea' refers to illnesses coded as 'diarrhea'; 'conjunctivitis' refers to illnesses coded as 'conjunctivitis'; 'other' refers to all other illness categories (residual).

TABLE 11

SUMMARIZING MISSING DATA - AN EXAMPLE
Kenya Project, Toddler Food Intake File

We have food intake records for 135 unique toddler ID's. Since the ID includes a code for sex, a scan was made for ID's that were identical except for the sex code, and the 16 duplicates were changed to the correct sex. Of the 119 toddlers, all but one had a birthdate on a file of toddler birthdates sent to us by Kenya Project. For these 118 toddlers, a "window" of 17 to 31 months of age was defined, and calander months with at least one food intake record were tabulated as follows:

88 toddlers had 12 or more continuous months of food intake.

6 had 12 or more months, but one month was skipped.

3 had 11 continuous months of food intake.

1 had 11 months but 2 months were skipped.

6 had 10 months of continuous food intake.

9 had less than 10 months (either 9, 8, 5, 2, or 1 month).

5 had no food intake observations (in the defined window).

Figure 1

EGYPT FOOD INTAKE (KCALS)

(See Table 8)

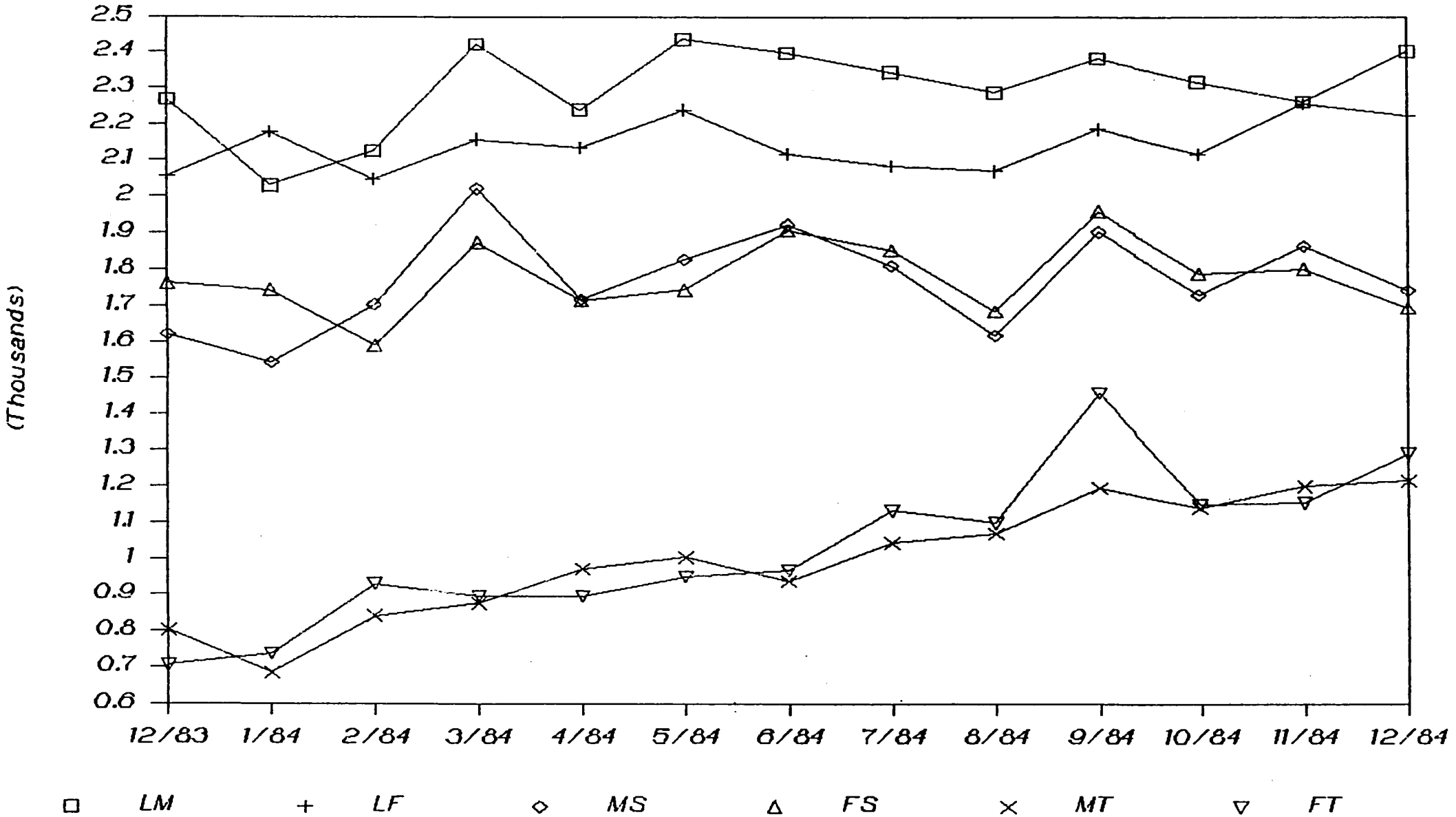
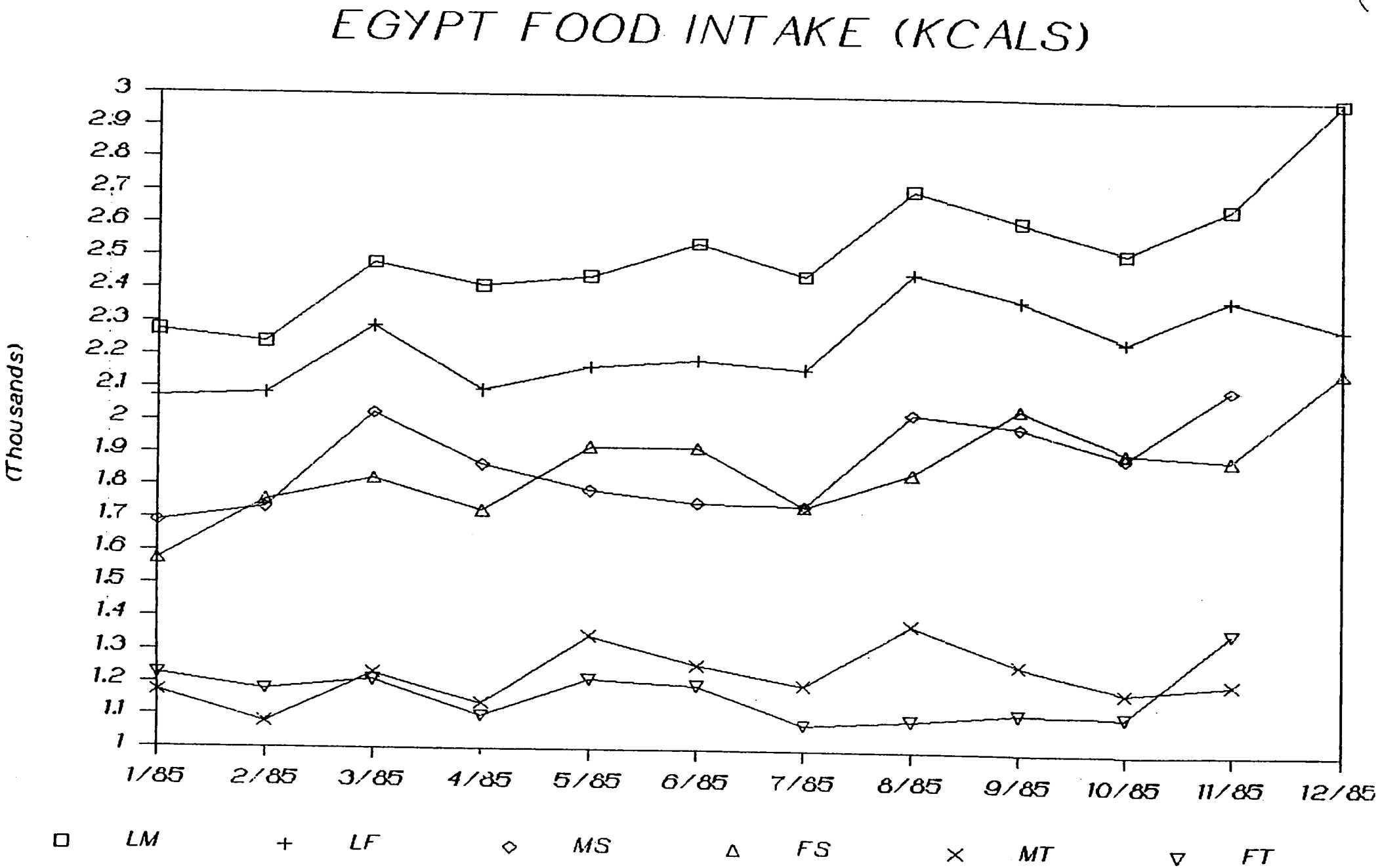
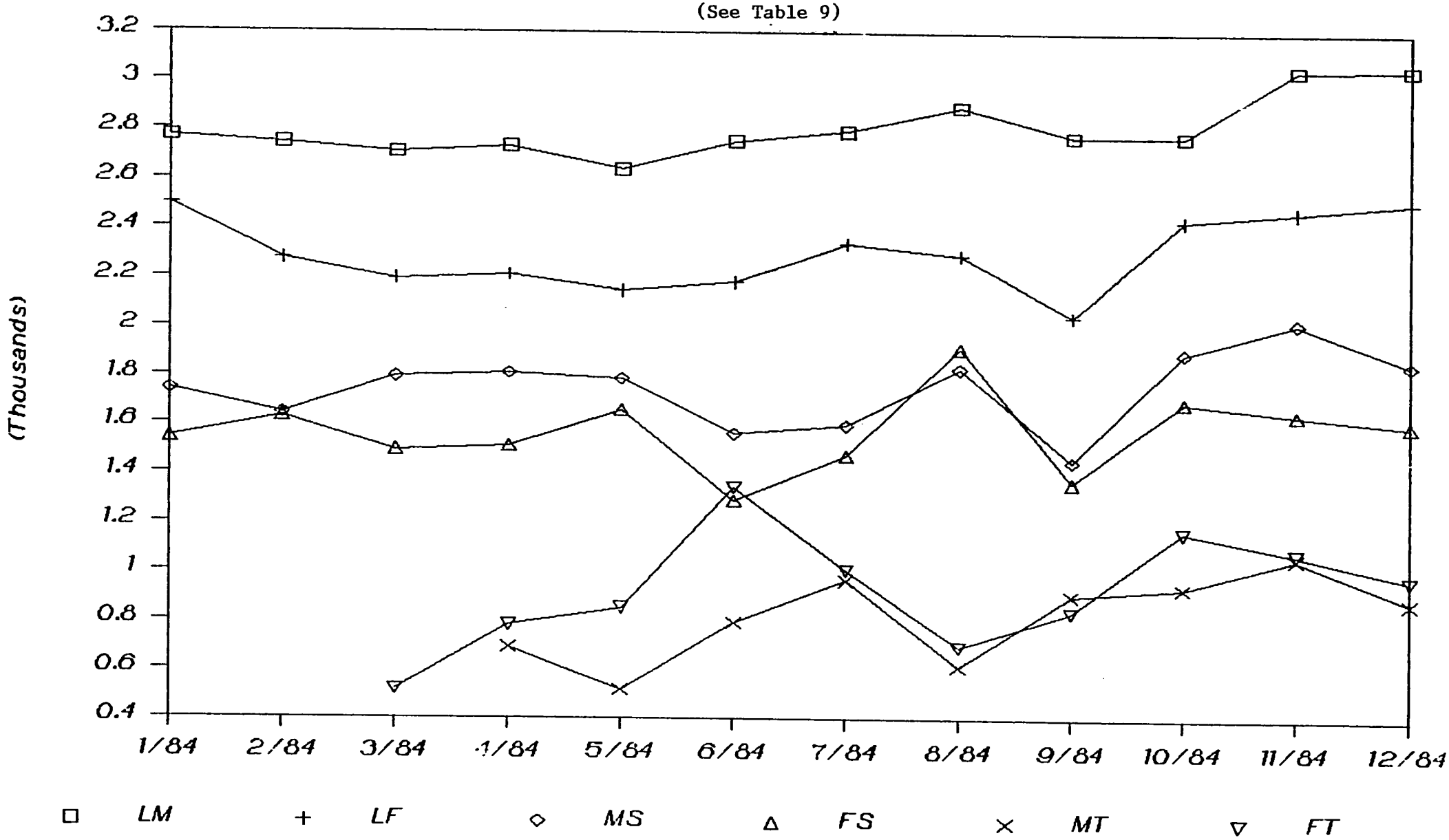


Figure 1 (cont.)



MEXICO FOOD INTAKE (KCALS)

(See Table 9)



MEXICO FOOD INTAKE (KCALS)

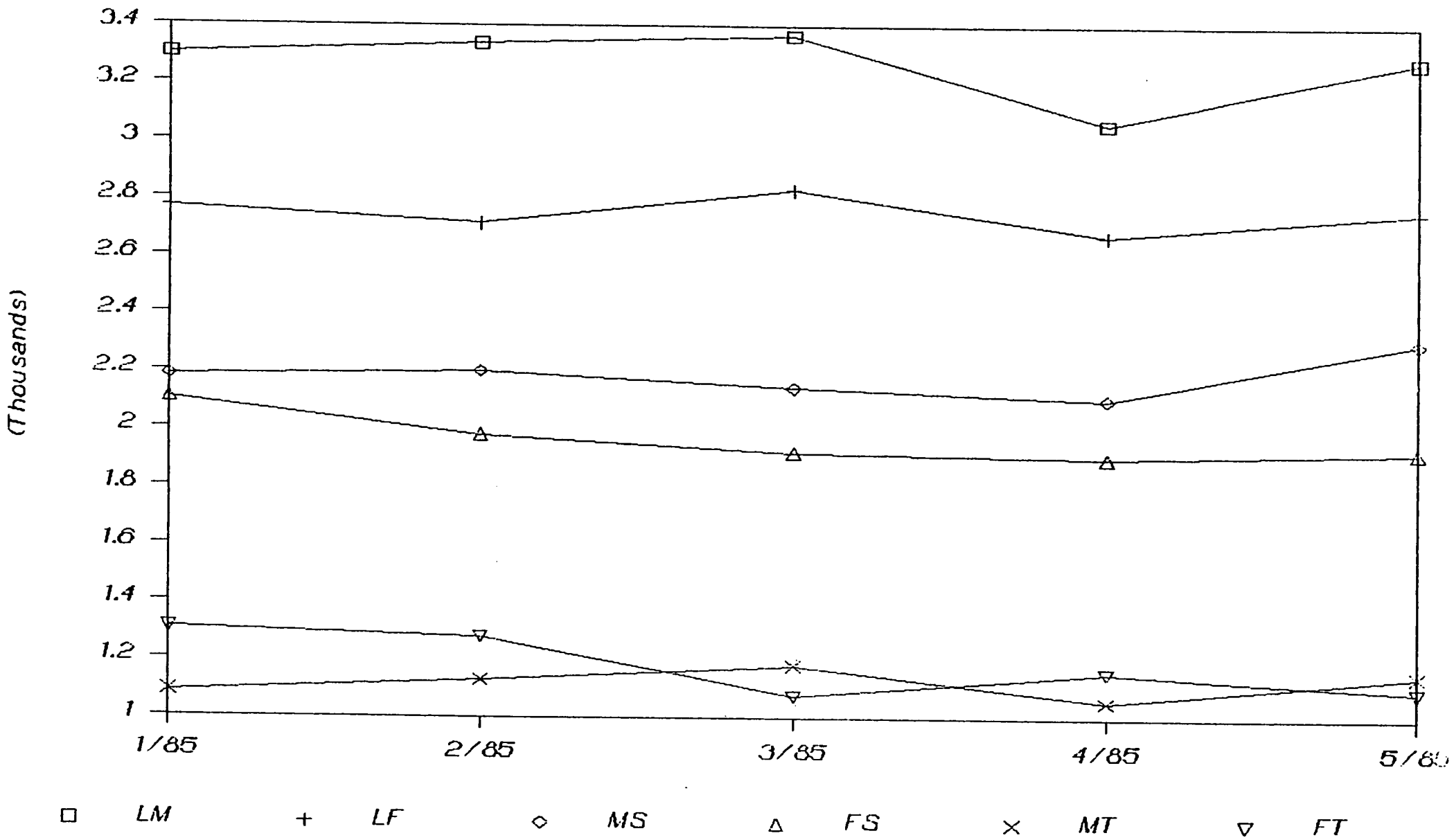


Figure 3

KENYA FOOD INTAKE (KCALS)

(See Table 10)

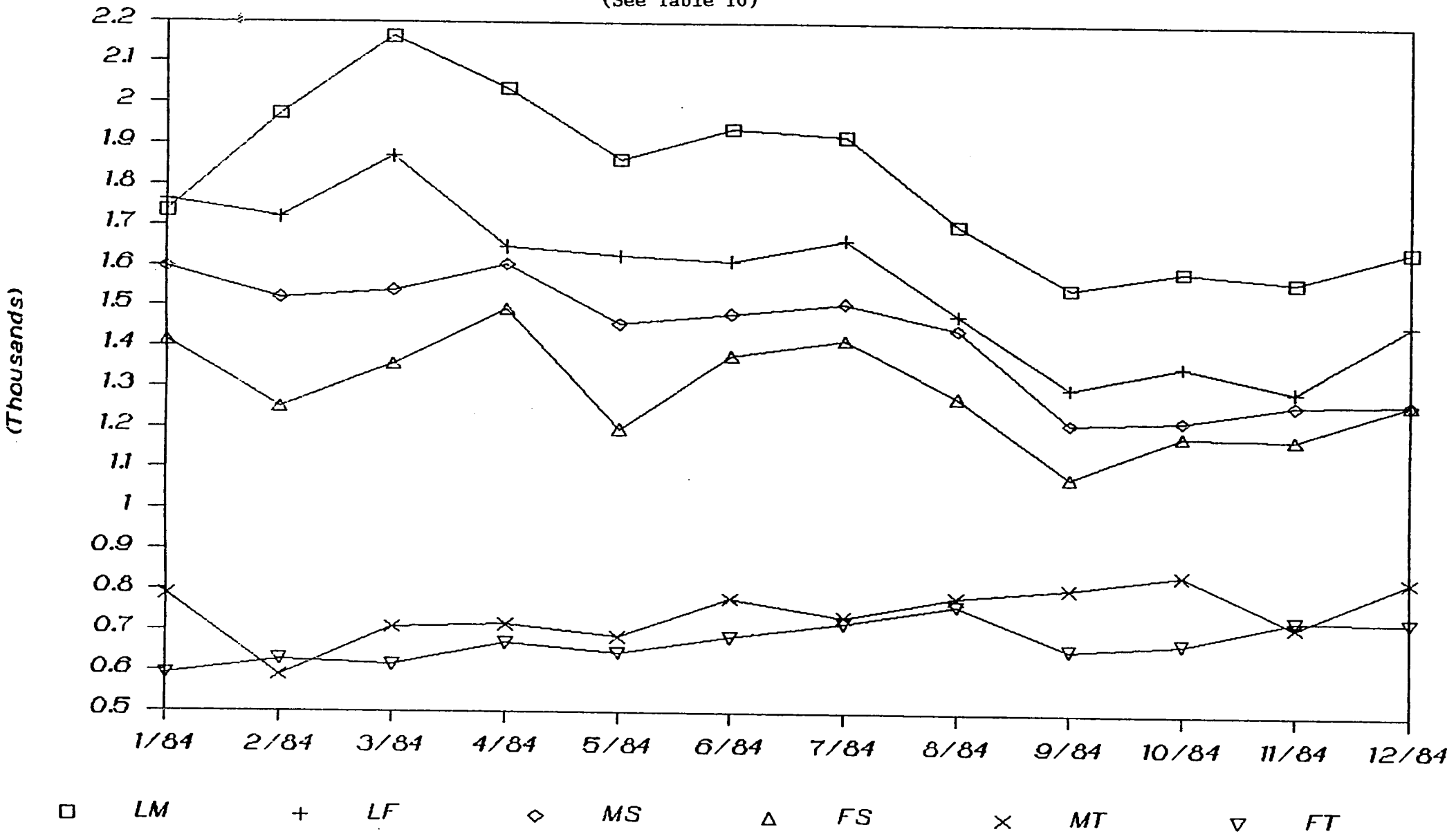
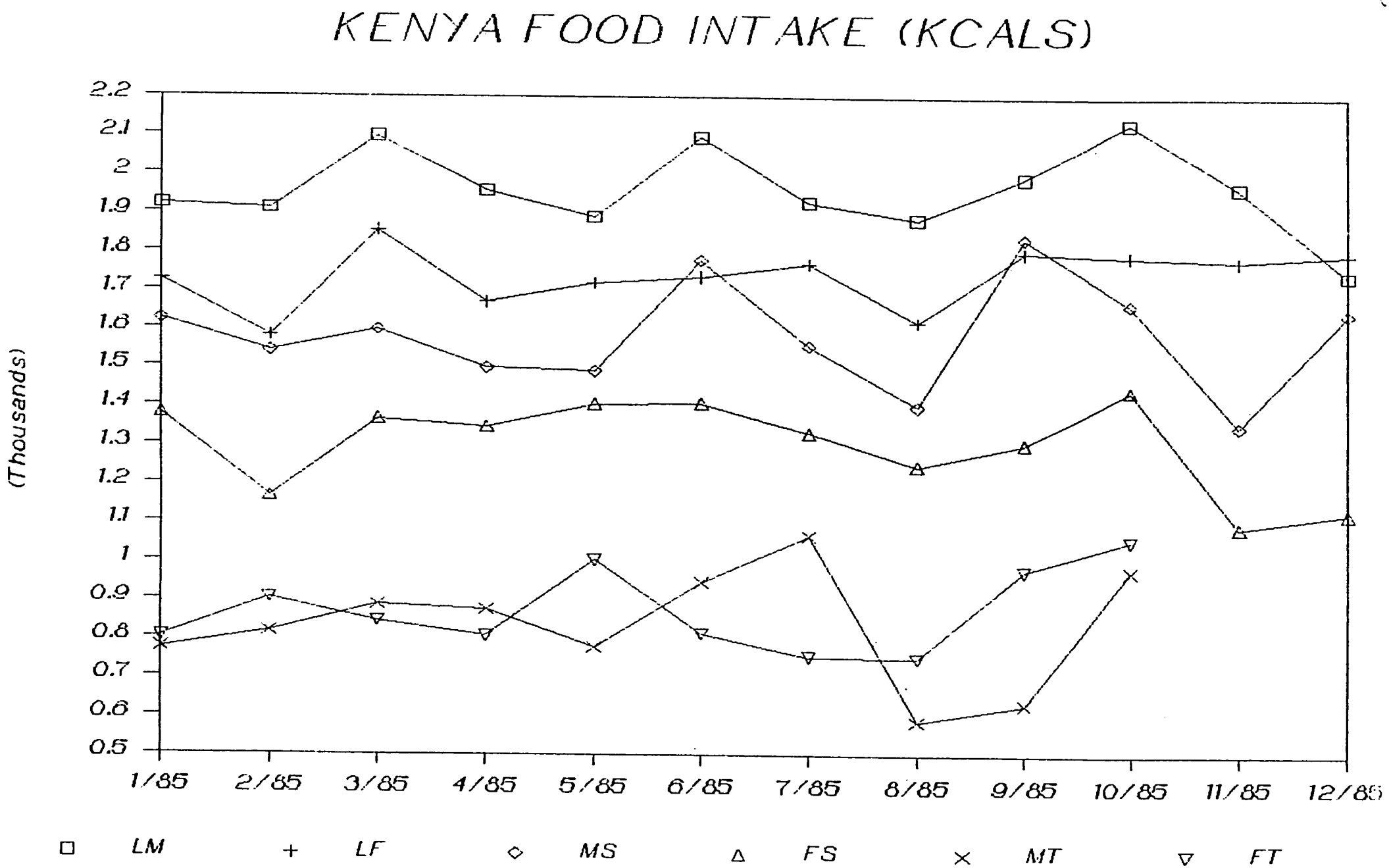


Figure 3 (cont.)



APPENDIX I

Dr. Lindsay Allen's SCB Meeting Minutes
Los Angeles (2/24-2/28/86)

RECEIVED JUL 2 3 1986

MINUTES OF SCB MEETING
Feb. 24-28, 1986
University of California, Los Angeles

Principal Investigators Present:

L. H. Allen, Chair
N. Bwibo
A. Chavez
O. Galal
G. Harrison, Vice Chair
N. Jerome
A. Kirksey
C. Neumann
G. Pelto

Management Entity Participants

G. Beaton
D. Calloway (February 27,28)
H. Horan

Monday, February 24

The opening discussion concerned a review of the agenda and a decision that the primary emphasis of the meeting should be on the definition of critical analyses. Dr. Beaton suggested that, in view of the limited time resources, it was important to create priorities among the various questions that ultimately will be addressed, that is, to decide on the key questions that could be reasonably explained in a limited time.

The original hypotheses were reviewed and it was noted that each of these are phrased in the general form:

"What is the effect of energy intake on _____?"

Given that the dependent variables (functional outcomes) are all determined by a multiplicity of factors, a concensus was reached that a more useful phrasing of the hypotheses from an analytic perspective is:

"What amount of variance in _____(eg. morbidity) is explained by food intake?"

With this clarification, we then turned to a review of the data base, area by area, in order to arrive at some conclusions about priorities. To assist the discussion, each area was assessed in relation to a series of criteria:

1) The value of the analysis for immediate policy

application, from host country perspectives

2) The value of the analysis as a contribution to basic theoretical-scientific interests

3) The strength of the data, assessed from the standpoint of research design

4) The quality of the data, as actually collected

5) The availability of the data (current and projected schedule).

Each of these criteria were assessed using a three point scale (3 = highest), for the areas of function, divided broadly into: morbidity, cognitive function, behavior (esp. maternal), growth, pregnancy outcome, lactation outcome and RMR. Table 1 summarizes the results of filling in the matrix.

In the course of discussion, a number of important issues were raised that should be part of the record of the meeting:

1) Dr. Bwibo pointed out that priorities for analysis in relation to policy depend heavily on <u>which</u> government groups one is addressing. For example, from the perspective of the Ministry of Health (in Kenya) morbidity, growth, and pregnancy would have high priority, whereas cognition RMR, and laboratory studies would be of low concern. On the other hand, the Ministry of Education would give the cognitive studies the highest priority. Moreover, the concept of "public" is complex: the "public" of reference for the public health community is very different from the "public" with reference to politicians.

2) Dr. Chavez expressed the view that behavioral consequences of marginal malnutrition are potentially the most important and interesting precisely from a policy perspective.

After considerable further discussion of the matrix, a concensus was reached that the primary analytic activities between March, 1986 and August, 1987 would be on:

1) Morbidity
2) Cognition
3) Growth

We would also attempt preliminary analysis in the area of behavior (especially maternal) and reproduction. However, it was

generally agreed that additional funding would have to be sought to complete the analyses of these areas. Table 2 summarizes these decisions.

Further Discussion of Analytic Issues.

The following comments provide, in outline form, a summary of a general discussion of analytic issues concerning the growth data:

A. Approach to growth

Examine toddlers as separate groups, e.g.
i. no weight gain for 3 to 6 months
ii. no height gain for 3 to 6 months

Define "no growth" -- Suggestion that the definition include e.g. for weight, the day to day variation in weight (from the 7-day daily weighing exercise)

Examine different growth slopes of different children

Beaton queried:

1) How do we characterize growth in children over 12 months?

Deviation from expected linear pattern?

2) How do we define the food intake variable?

How do we relate fluctuations in weight and food intake?

What is the appropriate F.I. variable for toddlers?

Suggestions:

1) Use of parametric vs non-parametric approaches

2) Three month moving averages to smooth curves

3) How do we identify real differences?

B. Statistical Challenges

How to define the cost of growth of 50 kcal/day (cost of daily growth?) This would correlate better with minimum weight and maximum weight.

By the time a child reaches 18 months too much has happened

to its growth -- Attained status may be most important.

Look at attained growth at different periods.

Suggested approach:

Look at the 50th percentile of attained growth.

The energy cost of 1kg weight-gain per year is 3 gms per day or 5 cal per gm. The direct cost of 3 gms is 15 cal/day. Over a year, a bigger cost is to maintain the body. A 1kg shift in body size is 60 cal/day.

In any model use:

Cost of growth
Cost of infection or illness
Cost of activity

The change in energy intake is too small to be detected except over the entire year. The cost of infection may blur it in the opposite direction. What is the partitioning of available energy? Perturbations of the system may be the only way to answer such questions.

Can't use RMR in toddlers!

Tuesday, February 25, 1986, SCB Meeting, Los Angeles
G. Harrison

The potential agenda was reviewed and discussed. It was decided to devote most of the morning to the scientific issues of food composition and RMR, and the afternoon to budgetary issues in order that these discussions could take place in advance of the arrival of the IC members.

Nutrient Composition Table Issues

Kenya project: Medallion laboratories has done proximate analysis on 30 to 40 samples, including composite recipes and common foods. The project is using these data to supplement USDA and other data in order to develop a composite food table and a food dictionary. This data base will be computerized and will enable the inference of nutrient values as well as energy.

Medallion labs is not doing fiber analysis. Suzanne Murphy

41

explained that dietary fiber analyses were expensive, and the decision was made that crude fiber would not be terribly useful. Instead it was decided to use the USDA method of calculating calorie value, which adjusts energy value of carbohydrate for the fiber content of the food. A general discussion of the fiber issue ensued. Dr. Beaton indicated that all available tables except the British ones overestimate available carbohydrate because they don't consider fiber. USDA handles this issue by using variable energy values for individual foods based on fiber content. All projects should check to see if any of the data in their project-specific data bases comes from British sources. Dr. Chavez indicated that the Mexican tables incorporate fiber. Dr. Allen suggested that Suzanne Murphy should coordinate the efforts to produce a statement from each project regarding how fiber is being handled on the conversion of food data into energy. Dr. Murphy accepted this responsibility.

Dr. Neumann indicated that the Kenya project thought they were under-estimating alcohol intake due to the legal prohibition of home-brewed alcoholic beverages. They conducted a special survey on the subject, and now know that some of the men in the study consumed considerable beer on weekends. The question of what to do about this known error was discussed. The problem seems to be confined only to lead males and the project has a good idea which lead males are affected. The consensus was that the way in which this is handled will depend entirely on the particular analysis. Dr. Beaton pointed out that the problem is analagous to the problem with breast-fed toddlers in knowing that we have incomplete energy information on those individuals.

RMR/BMR Data Review and Overview

Dr. Jed Gardner presented an overview of the RMR equipment and procedures for the Egypt and Kenya projects and of the data from Kenya. He defined RMR as being taken at least two hours after ingestion of food, beverages or cigarettes and at least 6 to 8 hours after vigorous activity. Both Egypt and Kenya are using the Beckman metabolic chart for measurement of RMR. The core protocol calls for RMS measurements every 3 months for lead males, lead females, and target schoolers and at months 5 and 8 of pregnancy and at specified intervals during lactation. Kenya lost data for January 1984, September 1984, and August 1985 due to equipment problems. The protocol involved bringing the subject to the field clinic where a physical exam took place followed by at least 30 minutes of quiet rest. The last five minutes of the rest were used to apply and check the inflatable mask.

Dr. Gardner reviewed in some detail the methods which are used by the metabolic cart for calculation and prediction. For calculation the cart uses the lowest value of oxygen consumption. The data form used in the Kenya and Mexico project was also

reviewed. Both projects are reporting the data from 6 individual minutes, i.e., minutes 18, 19, and 20, and minutes 28, 29, and 30 of the test period. In order to compare these data with steady state periods, Dr. Gardner selected 60 data tapes for analysis. Most subjects did not reach a steady state on the first visit. The difference between calculations from the steady state option in the machine and the six minutes which we are reporting was not significant in almost all subjects, and there were no systematic differences between the 2 data sets.

Dr. Gardner presented data from Kenya plotted for all age groups over time showing a clear effect of the famine and recovery, parallel to the trend in food intake. The data on pregnant women were also discussed. The average weight gain during pregnancy was on the order of 5 kilograms or less. RMR increased in pregnancy for most women but probably less than expected.

The question of cross-validation of methodology between Egypt and Kenya was raised. There were several attempts to make direct comparisons on individuals traveling between the two sites, but most of these didn't work for various reasons. On Dr. Gardner, whose RMR has been measured in both sites, results have been well within 5%.

Dr. McCabe presented the Egypt project data. About 3% of the RQ values are above 1.1, and more are 1.0 - 1.1. A lengthy discussion ensued of what really is being measured in this instance. Various causes of such a result including subject anxiety and carbohydrate loading were disussed. Dr. Beaton pointed out that it is necessary to know whether it is the oxygen or the CO_2 component which accounts for the abberent values and raised the question of whether they can be interpreted. Dr. McCabe suggested that VO_2 may still be a valid measure if CO_2 is affected by carbohydrate loading. The entire issue needs further discussion and analysis.

Dr. Allen presented the data, protocol and methodology from Mexico, where they have used a Max-Plank respirometer rather than a metabolic cart. This method gives the volume of expired air and the percent of oxygen, from which the oxygen and expired air is calculated. Thus the Mexican data do not allow minute by minute analysis. They have done "near basal" rather the resting metabolic rates, attempting to have subjects fasting early in the morning. The clinical exam and anthropometry are done at the same time as the BMR. The Max-Plank apparatus gives slightly higher values for metabolic rate than does the Beckman cart; in addition the difference between resting and near-basal rates precludes direct comparison between Mexico's data and the other two countries. The Mexican RMR data looks broadly comparable to FAO data. No major trends have been seen yet with regard to seasonal variations in metabolic rate.

A discussion of budgetary issues followed. Dr Horan presented the budget situation from the ME perspective, and the ramification of various approaches were discussed.

Wednesday, February 26

The sequence of discussions of the SCB included:

1. Discussion of growth and morbidity variables followed by a joint meeting with the statistics group.

2. Marian Sigman's report of some preliminary analysis of Kenya data for toddlers including food intake, morbidity and cognition.

3. Joint meeting with the IC concerning budgetary matters for Periods V and VI.

I. Discussion of Growth and Morbidity Variables

The PI's and others who were in attendance at the meeting divided into two groups for the purpose of discussing in small groups the variables of (a) growth and (b) morbidity. Dr. Harrison chaired the group which discussed growth and Dr. Chavez, the group which focused on morbidity. Following the group meetings a joint meeting was held with the statisticians to share the group reports.

A. Dr. Chavez reported from the group that focused on morbidity. Morbidity was divided into simple and compound variables as follows:

   1. Simple variables:
      a. Number of acute episodes per month and per 6 month period.
      b. Days sick per month (include chronic days) and per 6 month period.
      c. Duration per episode (needs definition)

      Corrected variables:
      d. Days with fever (or a disease with a fever)
      e. Days with decreased activity.
      f. Days without disease in family - with child <6
                                - without child <6.

2. Compound variables:
   g. Morbidity index (number of acute episodes x severity x duration)

3. Indirect variables:
   h. Total sum of changes in weight: absolute (wt)

4. Intervening variables:
   i. Crowding (number of persons in household)
   j. Presence of children.
   k. Sanitation
   l. Standard of living (SES)
   m. Intervention
   n. Season
   o. Time

Overall comments: Do not include chronic variables, such as conjunctivitis or colds, since they will dilute acute episodes. If an episode is separated by more than two days with the subsequent episode being the same, they will be counted as separate episodes. Otherwise, they will be counted as the same episode.

Chronic illnesses can be obtained from medical histories. The length of illnesses in terms of clinical expectation can be determined.

A question for discussion was: Are illness diagnoses coded or are different symptoms reported i.e. measles or skin rash? Finally, a decision was reached that a report will be sent to ME within two weeks (March 15) regarding the methods for constructing illness episodes and illness diagnosis for morbidity. The group assigned to develop the methodology is: Drs. Chavez (Mexico), Neumann (Kenya) and Rittenbaugh (Egypt).

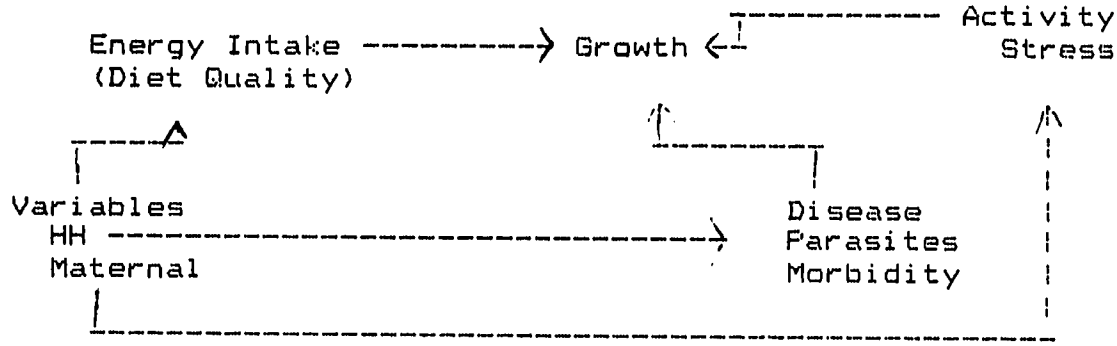
The questions to be answered by projects are:

1. Does morbidity condition growth?
2. Does food intake affect response to disease (severity and duration)?

B. Dr. Allen reported from the group which dealt with the anthropometry variable.

Household measures and how they may be related to growth were considered. The belief is the HH variables, if they affect growth, must do so through a biological mechanism.

Growth is affected by:



Growth may be called a functional outcome and predictors of growth may also be functional outcomes. The input variables to growth are energy intake and dietary quality. The detractor variables to energy intake are disease, activity and stress (affects utilization of energy). Adequacy of sampling in projects is minimal for activity and zero for stress.

The questions are:

Is the child growing?
Has the child changed weight over a period of time?
Is the child deviating from some model of growth?
Is the variability in food intake and in weight in the same child associated?
Is it simply noise in the measurement?
Can HH variables predict the status of growth?
Can a HH "suffering" scale be used?
Can a HH anthropometry score be developed?

Differences in velocity of growth are well established among the following three categories:

Infants, 0-6 months
Toddlers, 18-30 months
Schoolers, 7-9 years

The time interval needed for growth evaluation will be shorter for younger children than older The projected time intervals needed are: 1 month for infants, 3 months for toddlers and 1 year for schoolers.

Parents height may be used to control for biological variability in growth of children. Subjects may be categorized according to no growth with evidence of no growth defined as

	Time period (months)	Alt time period (months)
Infant	1	3
Toddler	3	6
Schooler	6	12

II. Report by Marian Sigman of Preliminary Analysis of Cognitive Data for Kenyan Toddlers

Dr. Sigman emphasized that the Kenya data which have been analyzed for interrelationships among food intake, morbidity and cognition are very preliminary and should be considered only as provisional findings. She described certain aspects of the Bayley developmental scale and the observational data (2h observations) which were used in the Kenya setting. She indicated that some data were collapsed for interaction observations to early and late periods as follows:

Toddler age, mo	Early Period age, mo	Late Period age, mo
18	15-17	18-20
24	18-23	21-23
30	24-29	24-29

The preliminary findings to date have indicated some influence of morbidity and less of food intake on the selected cognitive measures which have been examined.

III. Joint Meeting of IC and PIs

The SCB and IC members met jointly to discuss issues and problems associated with the Nutrition CRSP budget for periods V and VI. IC members requested the meeting with the PIs in order to allow the PIs to react to several budgetary recommendations before final action was taken by the IC.

Dr. Horan reported that IC had adopted the following recommendation concerning UCLA's anticipated overrun of $168,000 in Period V.

1. Attempt to resolve the overrun with resources available to UCLA and/or ME.
2. If it is not possible to resolve the problem from UCLA or ME resources, request necessary supplementary funds from AID.

Dr. Horan presented a summary of the plan approved by the IC for the UCLA cost overrun in Period-IV and the projected overrun of some $168,000 at the end of Period V. Dr. Horan indicated that ME is willing to release the balance of UCLA's funds (subject to release of all budgeted funds by AID) on the condition that funds allocated for data delivery are restricted to costs incurred after October 1, 1985. He indicated that ME would monitor the cost and delivery of the documented data. Dr. Horan also indicated that the Mexico project could release up to $50,000 and ME $82,000 - 86,500 to provide for minimum analysis of the Kenyan data in Period VI. This proposed course of action was acceptable to the PIs.

Dr. Horan summarized the budget levels proposed for Period VI as follows: Egypt $321,875, Mexico $336,093 - 338,593, Kenya, $132,000 - 139,000 and ME $583,457 - 587,957. These levels were accepted by the PIs.

A resolution proposed by Dr. Smith (Purdue) to the IC was presented to the PIs for discussion. The proposal was:

"In the event that a project component cannot complete its total work assignment within its authorized budget alternate funding sources will be explored with the following priorities-

1. Other funds available to the institution experiencing the shortfall
2. Funds from the ME as a contingency allotment
3. Direct request for funds through ME to AID
4. A request to IC for recommendation of reallotment of overall CRSP funds."

The proposal was discussed and was found to be unacceptable to the PIs.

At the request of Mr. Edwards, the plan and timetable for data analyses were presented to the IC. The PIs agreed that they could accomplish the objectives for data analysis for CRSP at least at a minimum level by August 1987 except for the area of behavior.

Thursday, February 27, 1986

Doris Calloway joined the SCB meeting in the morning, as did Martin Forman. The topics of discussion, in the order in which they were taken up, are as follows:

1. The Methods Manual. The feasibility of the previous arrangement was reaffirmed: each project will write up all portions of their project first. Responsibilities for reviewing these materials will be divided, following the lines agreed on at the previous meeting. Norge Jerome will take responsibility for reviewing the description of the data management component.

2. Cross Project Analysis. In reviewing the work of the SCB on Monday, George Beaton pointed out the change in the analytic work from the testing of specific hypotheses to an examination of areas of function. He reaffirmed the decision that all analyses could not be done in parallel, that there would be no "pooled" analysis of data and no separate interproject analysis. The three projects would examine the same issues in concert.

George pointed out that the first step was parallel, cross tabulation tables, an exercise that was important but too slow as a general mode for the future. Rather, what is feasible is work in concert on defined questions, with shared approaches.

In the coming months each project will put full investment in morbidity and cognition. George Beaton suggested that it is not clear that the statistical analystic models will be identical for each project and it is probably not practical to have programs shared across projects.

Finally, Doris emphasized the importance of focussing on longitudinal data, while George pointed out that the specialists in particular areas (such as psychology) will not be very helpful for identifying appropriate statistical models for longitudinal analysis since this is rarely done.

3. Analysis Activities in the Host Countries.

a) In Kenya there is great interest in carrying out data analysis of project results. The main constraint is money for computer facilities and staff time.

b) In Mexico a data analysis unit for the project is already in place doing data management activities. This unit will shift to analysis as data management responsibilities are completed.

c) In Egypt the limiting resource at the institute is computer resources. The project intends to build an analytic counterpart to the unit in the US.

4. The Policy Committee. Under Dr. Forman's direction, there was a discussion about how project data can be utilized for policy development. The host country PIs all emphasized the current pressures on them to provide descriptive data on results.

In Egypt there will be a seminar in 1986 sponsored by the National Academy on project results. Egypt has a large sum from AID through the Child Survival Project and is in the process of setting up a five year project. A CRSP investigator is involved in the advisory committee.

In Kenya several groups have expressed a "need to know" project results, including Meals for Millions. In Mexico, project results can be directly disseminated to action programs within the Institute of Nutrition, which is responsible for 80 - 90% of all national programs in primary health related to nutrition.

5. Data Management. The data managers met with the SCB to present for discussion the issues raised in their meeting. The data managers pointed out that, given the limited staff

resources, there is, in effect, conflict between analytic needs and archiving needs. They suggested the value of a more formal allocation of resources between the two components.

Suzanne Murphy emphasized that the conflict is not trivial, which led to a discussion about when documentation can be considered "adequate." What emerged from this was a clearer understanding of the difficulty of specifying documentary adequacy. Given the time and resources, there will, ultimately, have to be compromises with what the data managers would consider truly adequate documentation (with full flagging of all data).

6. Consumer Units. George Beaton led the discussion, beginning by reviewing various options, which are as follows:

1) total predicted (estimated) household BMR
2) computed energy requirements (based on normative standards and weight)
3) computed energy requirements based on age and sex
4) a standard "consumer unit" in which an adult male = 1.0, an adult female = .75, etc. based on normative energy requirements

After considerable discussion, Doris Calloway proposed that the value be calculated on a day by day (but not meal by meal) basis. The denominator is BMR and the unit is all of the people reported to be present on the day of measurement.

Friday, February 28

Cross Project Analysis

The morning session began with a discussion about cross-project analysis. Dr. Calloway pointed out that ME has the responsibility for this interpretive task and that the participation of the PIs is welcomed and encouraged.

Dr. Beaton described the parallel cross-project analyses as occuring on two levels:

a) the effort, to date, which has been essentially focussed on identification of problematic data

b) the statistical-analytic activities of hypothesis testing.

The communication chain for both levels is now in place. However, the interpretive activities associated with the second level will take considerable collective effort. Dr. Beaton noted that the Bellagio conference format would be the best mechanism

to achieve this.

Publication Policy

Dr. Calloway led the discussion concerning publication policy, which addressed the following:

1. ME will send one copy of project data tapes back to the host country institutions and the US PIs in August, 1987. The SCB agreed that copies of data tapes from all three projects should be sent to each project, so that all PIs will have access to all data during the one year (8/87 - 8/88) before the tapes are made available to the wider public.

2. Publication policy is only enforceable to August, 1987, at which time the CRSP as a formal entity will cease to exist. After that time, the policy is based wholly on ethical guidelines.

3. After August, 1987 any corrections to data sets will have to be documented, by memo, directly among PIs.

4. Reviewing current policy, Dr. Beaton made the following points.

a) Concerning release of data, ME won't release any data without permission of the project representative PI, to 8/87. Projects won't release their own data without SCB review and approval. This applies to tapes, data summaries and papers.

b) SCB will review requests for exceptions

c) descriptions of methodology as releasable

d) no project will publish on the basic hypotheses until relevant data sets all are certified complete by the project concerned.

Dr. Beaton suggested that the only position the SCB can adopt is that for within project data, the data set has to be complete. For example, within a project, the psychologists can publish methodological papers but they cannot publish substantive results, until the PIs certify that the relevant data sets are complete.

Dr. Neumann raised the matter of their desire to give project data to an NGO that is beginning a program in Kenya immediately and want to see their data for the purpose of program planning. They are going to have a seminar in October, for which these data are requested.

Dr. Horan suggested that to speed the process of approval, ME would take responsibility as follows:

1. To mail a request within 3 days of receipt.
2. To include a cover letter with the statement that no answer within 30 days would be taken as approval.
3. The letter will be sent to the US PI representative, who has the responsibility of communicating with the other project PIs, unless there is a blanket approval agreement.
4. The process of SCB approval will occur when there is any release of data (tapes, tables, publications) until a project certifies that it has a complete and clean data set.
5. ME will not release any data without project approval until 8/87.

The exception to this general statement is that a project may release data to a national or local planning group, for the purpose of proposal writing or for job interviews. As a courtesy, ME should be informed when this occurs.

With respect to Phase I data, any papers that do not test the basic relationships among nutritional status and function must be submitted for SCB approval, until 8/87. Publications that test these basic relationships cannot be put forward until the relevant data sets are certified as "clean and complete" by all of the PIs of the project.

In the case of conflict within a project, the SCB is the appropriate forum for action, not ME, according to Dr. Horan.

Dr. Galal pointed out that in Egypt, findings reported in thesis are not regarded as "counting" toward professional advancement, so students tend not to publish their analyses. However, their professors do. With respect to CRSP data, there are already several theses completed under the sponsorship of senior scientists. In this case, it is now too late to control the publication of these results.

Finally, with respect to publications, the matter of using CRSP funding to offset publication costs, after 8/87, was discussed. Dr. Calloway reported that ME planned to cover the cost of duplicating tapes and final reports, both of individual projects and the joint, interpretive effort. She suggested that a monograph might be the best vehicle for disseminating results.

Since a no-cost extension beyond 8/87 is highly unlikely, it is clear that there is a serious problem with dissemination of results through professional journals to the world-wide scienctific community.

Future Funding and the Future of the CRSP

52

Dr. Calloway raised, for discussion, the issue of how future analyses will be completed once CRSP funding is exhausted. A central question is whether a collective effort, in the form of a not-for-profit corporation, can be more effective and successful in securing funding than individual project proposals. The group response to these questions was quite tentative and mixed. The main problem, it appears, is that no one has enough information at this point to feel confident about which avenue is more productive to pursue at this time. Of central concern is the matter of cost. Can there be economy of scale or will it be less expensive to pursue smaller-scale, individual project, funding?

With respect to future activities:

1. Dr. Galal reported that the Ford Foundation has given the Institute of Nutrition microcomputers and the analytic group at the Institute would like to use Phase I data for a workshop whose aim is to answer some questions for the child survival project.

In a broader perspective, the presence of the CRSP has furthered the development of other projects, including proposals now pending with funding agencies to develop Egyptian food composition tables and to develop a project on dietary control of diarrhea. Dr. Galal reports that the CRSP has been particularly helpful in facilitating real integration of biological and social sciences, as well as stimulating discussion about nutrition and health issues.

2. Dr. Forman, who joined the SCB discussions on February 28th, reviewed other AID actions in the area of nutrition. He stressed the significance of Child Survival programs and suggested we look for ways in which to link up CRSP activities to such programs.

3. Dr. Forman requested that the projects prepare reports on how the CRSP has strengthened local capabilities, including an inventory of training, equipment and other institution building activities. This report should include the impact of the project on the local community.

Future Meetings

1. There will be a meeting of the statisticians in late June at Purdue.

2. A late fall meeting, probably in Berkeley, will include some members of the EEP. A primary purpose of that meeting will be to review progress on data analysis, with an emphasis on comparative results.

TABLE 1

FUNCTION	POLICY What a country needs (3 point scale)	Theoretical Scientific Considerations (P.I.)	Design (P.I.)	Known Quality	Data Availability
MORBIDITY	Egypt (Galal) 3 Mex (Chavez) 2 Kenya (Bwibo) 3 (average 2.8)	3 (Mex. 2)	2	Mex Egp Ken Overall "good"	All projects have some dat now available
COGNITIVE	2 2 1 (2 for Edu. Min.) (average 1.7)	3	3	3 to 3+	Mex: end of Apri or May '86 Kenya: May '86 Egypt: July '86
BEHAVIOR (Maternal)	3 3 3 (average 3--) Unique to CRSP	3	3	all 2.5	Mex: needs time code - End summer Kenya: Spring - Summer Egypt: behind in ing/August
GROWTH	3 2 3 (average 2.8)	3 (Mex. 1)	2	all 3	All projects have data available
PREGNANCY OUTCOME	3 2 3 (average 2.8)	3 (Mex. 1)	2	all 3	"
LACTATION OUTCOME qual/quant.	3 3 3 (average 3)	3 (Mex. 1)	2	all 3	" Mex. lab aspect Aug. '86
RMR	all 2-3 (average 2.5)	3	2-3	Mex Egp Ken all good	Mex. ? Kenya May Egypt ?

54

TABLE 2

Priorities for Analyses and Time Frame

Functional Area	Planning for Analyses	Preliminary Analyses	Completion of Analys
MORBIDITY	2/86	6/86	6/87
GROWTH	2/86	6/86	6/87
COGNITIVE	6/86	9/86	6/87
BEHAVIORAL AREAS *	9/86	1/87	?
REPRODUCTIVE	8/86	1/87	6/87

* Hopefully obtain funding for full analyses for Cognitive/Behavioral after 1987.

APPENDIX II

Dr. Doris Calloway's Memo on Descriptive Statistics
Berkeley (4/28/86)

28 April 1986

To: PI's, statisticians, data managers, ME associates

From: Doris Calloway

Several long flights have provided me with an opportunity for uninterrupted thinking about the CRSP. Since variables describing food consumption will be critically important in all analyses, I have outlined some suggestions about possible approaches to descriptive statistics for these variables. They include various ways of looking at mean intakes, variability of intakes, and relationships of intakes to illness, anthropometry, and pregnancy/lactation status. I also have made a few suggestions about RMR/BMR descriptive statistics. It may be appropriate to approach these questions first for a specific target type, and then extend the analyses to all individuals. I think a unified approach to these important questions will be in keeping with the CRSP objectives, and will benefit us all.

I hope you will have time to discuss these suggestions and questions within your projects so that agreement on analytical strategies can be reached at the June meeting of the statisticians, in preparation for the Fall, 1986 round-up.

57

I. Food Consumption

A. Description of individual intake data (2 days/subject/month)

1. Calculate energy (and protein?) intake

a. By time periods:

-12 months mean/subject for all values available for a subject; record number of values per subject.
-quarterly (3-month blocks, 4 values per subject.
-every 3 months moving averages [complete set would have 10 values per subject (except pregnant or lactating where 4 values each during pregnancy and lactation)]
-"season" (to be defined by each project)

b. The above mean intakes expressed as:
-mean per diem
-mean per kilogram body weight (Note: This could be done as weight recorded on day one of relevant period or date closest to it, mean weight for the total period, or midpoint weight.)
-mean divided by BMR, calculated using FAO formula (based on age, sex, weight) and actual RMR/BMR if available.

c. Stratify by target type, and within target type according to intake in relation to standard, or multiples of BMR.

2. Examine morbidity records: by subject; by target type; by subgroups stratified by intake in relation to a standard (as in 1c above)

a. Note proportion of days of the intake record which coincide with:

- days of recorded illness (be sure that regularly scheduled intake was recorded on ill days; if not, then intake is biased);
-recovery days, the period 10 days after the end of an illness;
-prodromal days, in the period 2 days before illness;

(Query: how many days are there typically between illness episodes? is it appropriate to assume at least

12 days between illnesses - 2 prodromal and 10 recovery?)

(Second query: would aggregates of recorded-sick days and of NOT-recorded-sick (i.e., "well") days be sufficient, rather than try to define "recovery" and "prodromal"? is "recovery" definable?)

b. Compare proportion of days ill in the following time blocks: yearly, quarterly, moving 3 month averages, seasonally (as in 1.a., above).

c. Compare proportion of "recovery" days (or NOT-recorded-sick days) in the time blocks in 1.a., above.

B. Identify and classify subjects according to variability of within-person intake.

-What characterizes subjects of high and low variability?

-Principal component or cluster analysis? Possible variables: socioeconomic; village/block; sanitation; number in household; birth order; presence of lead male 7 days per week, 2 days per week, never; morbidity; height/age; weight/height; household food/household BMR; number of eating events per day; sex.

-Within households, which target subject's intake is most/least variable?

C. Identify and classify subjects according to mean intake-

-What characterizes subjects with high and low intakes? Proceed as in B above.
-Are the same subjects classified as high (or low) in all time blocks?

D. Intake and illness

-Is intake decreased during illnesses of different severity? Increased during "recovery"?

-If intake is decreased during illness of subject, does household food decrease? If not, is intake of any other target subject changed? Who benefits from a subject's diminished appetite?

E. Intake and anthropometric characteristics

-By target type, what is the intake of subjects who gain/lose/maintain weight during lagged 3-month periods (10), quarters (4), seasons (maybe Ramadan in Egypt?), year? Expressed as kcal/day, kcal/kg, kcal/BMR (calc) and BMR/RMR (measured).

-Compare 12-month intake of toddlers and schoolers with gain in height. (Smaller time blocks are not appropriate for height analyses)

-Compare maternal lactational intake (intake divided by BMR and per kilogram) during 6 months with 6 month gain in weight and length of infant. Look at this question with data sets stratified by birth weight; maternal 6-month gain in weight during pregnancy; maternal skin-fold thickness at delivery; change in maternal skin-fold thickness (from delivery until 7th month).

F. Maternal Food Intake

a. Is there an association between maternal intake during pregnancy and during lactation?

-between maternal weight gain during pregnancy and intake during lactation?

-between maternal weight gain during pregnancy and intake during pregnancy?

b. If there are enough non-pregnant, non-lactating women, (NPNL) in the data set, is intake of pregnant women greater than NPNLs? lactating women's intake > NPNL? Do these relationships differ seasonally (as in Gambia)?

G. Household intake

-Is there an association between household food intake and target's intakes?
-Does household intake vary with household size, composition, etc.?
-Does household intake vary with lead male/female cognitive test scores, occupation, SES? Morbidity?

H. Diet quality

a. Does percentage of energy from protein differ between target groups? and households? with illness? recovery?

b. Micronutrients--do they vary with energy intake? If so, is the lowest intake range "probably adequate"?

II. BMR/RMR

A. Does measured BMR/RMR vary with age in adult targets? with height? with weight?

B. Does measured BMR/RMR vary with energy intake (per diem, per kg) within sex/age groups?

C. Does measured BMR/RMR differ during pregnancy (per diem, per kg)? Seasonally? With illness?

D. Mexico: do your measured BMR data coincide with FAO calculations?

APPENDIX III

Dr. George Beaton's Minutes of Meetings with UCLA Staff
Los Angeles (5/21-5/22/86)

Notes from meetings of G.H. Beaton with UCLA staff (Neumann, Afifi, Gardner, Ben Browdy, Coulson, Sigman, and returned field staff). May 21-22, 1986.

1. The meetings provided an excellent opportunity for informal collegial discussions of analytical activities and issues that, I think, was productive for all parties.

2. The main topic was RMR (see below), but a number of other items were touched upon or discussed.

3. Morbidity Data

o Dr. Neumann suggests that Berkeley defer efforts to work with the Kenya data set pending completion of review and revision of data files. There are differences introduced by a change of forms, of personnel, and of coding. These are now being corrected and new files will be sent to Berkeley as soon as available. In the meantime, there is little merit in either working with, or querying for correction, the existing file.

o The Kenya Team is developing and testing a new classification scheme for field morbidity data. (Primary Kenya data includes the field diagnosis rather than only symptoms--this is probably a different structure from either Egypt or Mexico.)

o The classification will operate by either single disease Type or by classes of disease (grouped diagnoses). In the groupings, diseases are being selected for their probable/possible relationship to food intake and/or nutritional status (from literature review). A number of disease conditions that seem almost universal but of little or no nutritional or health significance are grouped in one class. (See below.)

o A "severity" scale is also under development. This will yield a numeric score based on the clinical severity of the disease, reported effect on appetite and food intake, and reported effect on activity, weight, and duration. Duration is discounted (lower weighting) after two weeks and again after one month. As noted, the issues involved in applying this are now being identified and addressed.

o Overall, morbidity records are also being classified in terms of
- well days
- sick days- attributed to mild diseases
  - attributed to significant diseases
  - attributed to specific illnesses (individual or categorical)

o The approaches being attempted look interesting. It is premature to offer evaluative comment--some of the categorically scaled episodes are being plotted on the longitudinal weight/height/intake/etc. charts. There are suggestions that there may be associations with linear growth.

o Dr. Neumann expressed the hope that she could provide a preliminary report on the approach for discussion at the statistics meeting.

4. Cognition/Psychology

o It appears from brief conversation that the initial "disciplinary" analysis of the toddler psychology data is almost complete. In process at the moment is an editing down of the observational data to bring it to a standard time base. (Because of disruptions/target child going to sleep/-etc., the recorded time from beginning to end may exceed the protocol time but the record has periods of non-observation in the middle. They are now being reviewed and standardized to constant period of observation times.)

o Little has been done yet (?) on schooler or adult data.

o Dr. Neumann is soliciting funds from a possible donor to support schooler cognitive data.

5. SES Scaling

o Work is beginning on new SES scales with input from returned field staff (and, in turn, based also on inputs from community leaders).

6. Validation Study - Toddler Weighing

o The results of the weighing-reweighing validations were presented briefly (they are included in the project report). The results are quite similar to those from the Egypt trial with observer effects being very small or small, but significant, within person effect. The project will also identify validation data from the field (i.e. when a quality control observer was present) and derive estimates there.

7. RMR

o The bulk of the time was spent on these measures. The data set is now complete. Kenya project staff had not appreciated the magnitude and potential significance of the "high RQ" issue. A major reason was that they had to undertake intensive examination of the first 3-4 months of field data and found no problems, other than an apparently low RMR compared to predicted. There was surprise about the analyses of RQ performed by Berkeley and a realization of the import of establishing the basis of these high values before any interpretation could be made. There was full and free discussion of possible sources of "instrumental error", "operator error", and "subject preconditioning", without clear resolution of the issue--but agreement that it must be pursued.

o Ben demonstrated, with excellence, the advantage of ready access to computer files. As issues or questions arose, Ben was able, in a period of 10-15 minutes, to generate and present for discussion cross-tab printouts that permitted examination of the postulate. This was an outstanding achievement and I sincerely congratulate the Kenya project for developing this capability.

o It appears that they had done detailed analyses of the first 4 months of data and satisfied themselves that all was well <u>except</u> that RMR's were low in comparison to predicted values. In retrospect, the values were good at that time and the drift in RQ began the two months later.

Date		Mean RQ_1	Mean RQ_2	"Best"RQ	VO_2	VCO_2	REE
1984	Jan						
	Feb	0.90	0.91	0.91	173	158	
	Mar	0.89	0.86	0.87	165	145	
	Apr	0.90	0.90	0.90	145	130	
	May	0.82	0.83	0.82	155	127	
	June	0.87	0.87	0.86	171	149	
	July	0.96	0.97	0.99	149	142	
	Aug	1.05	1.06	1.03	135	138	
	Sept		(machine computer out)				
	Oct	1.08	1.08	1.06	128	137	
	Nov	1.04	1.04	1.02	131	137	
	Dec	1.05	1.06	1.06	141	140	
1985	Jan	0.85*	0.86*	0.85	166	144*	
	Feb	0.85*	0.85*	0.85	190	161*	
	Mar	0.85*	0.85*	0.85	158	133*	
	Apr	0.89*	0.89*	0.88	172	152*	
	May	1.02	1.02	**	164	167	
	June	1.02	1.03		178	182	
	July	1.07	1.05		183	193	
	Aug	0.98	0.98		186	185	
	Sept	0.95	0.94		182	172	
	Oct	0.95	0.93		167	159	
	Nov	0.95	0.94		180	169	
	Dec	0.98	0.96		156	151	

* CO_2 analyzes out; RQ = 0.85 assumed and VCO_2 calculated.
** Data = 1 or above were entered as 0.99.- i.e. data set attenuated and unusable beyond this date.

The mean VO_2 and VCO_2 were examined by categories of RQ. This table must be interpreted with care since, after April 1985, all RQ's above 0.99 were entered as 0.99. This means that the category 0.9 - 1.0 also includes some RQ's of 1 or more!

Recorded RQ	n	VO_2	VCO_2
<0.7	21	144.7	140.6
0.7-0.8	23	159.3	123.0
0.8-0.9*	298	170.0*	144.6*
0.9-1.0**	257	164.5**	163.2**
1.0-1.1	34	134.7	140.6
1.1-1.2	33	126.9	145.7
1.2-1.3	22	117.5	150.3
1.3 +	19	139.0	150.5

* Includes values with assumed RQ = 0.85 and calculated CO_2
** Includes some RQ's above 0.99 (last part of data set).

What is particularly striking about the data set is the <u>appearance</u> that O_2 is driving the RQ more than CO_2. However, because of anomalies in the data set, this should be recombined with calculated RQ's. Nevertheless, if the O_2 is the determinant variable, then one must at least suspect that the O_2 analysis was

malfunctioning--but why only for some cases remains unknown. If all O_2 values were underestimated there are at least two critical implications:

1) the whole distribution of RQ's is biased upward
2) the RMR of Kenyans has been <u>underestimated</u>.

The most important implication is that the whole data set might have to be discarded or unreliable and hence uninterpretable. That is, the apparent fall in VO_2 and in RMR in association with the famine <u>could</u> be artefactual--certainly the RQ's were out of line at the time!

A check on the VO_2 and some other parameters will be available for the final period since other equipment was moved into place for exercise physiology studies. These data have not yet been worked up. However, it was reported that the BMR estimates with the new equipment were reasonably consistent with the predicted BMR while the Beckman continue to give estimates of BMR appreciably below BMR. This might again suggest that the Beckman was underestimating O_2.

There is urgency in having these data comparisons undertaken. All possible parameter comparisons should be made.

The Kenya project might be urged to recover a sample of the calibration gas and independently validate its composition to see if this could be the source of a problem.

Other data sets were available for examination. These are summarized in the next table.

Date	Mean Time of Day	Mean Oral Temp °C	Mean Room Temp °C	Mean Barometric Pressure	Humidity
1984 Feb	11.6	37.2	--	611.9	--
Mar	11.8	37.1	--	612.1	--
Apr	12.2	37.2	--	612.5	--
May	10.6	37.0	22.6	613.1	66.9
June	11.1	36.9	22.2	612.9	65.1
July	11.5	36.7	21.7	613.2	73.2
Aug	12.4	36.7	20.4	613.3	79.2
Sept	--	--	--		--
Oct	12.8	36.8	22.9	613.1	75.5
Nov	12.8	37.0	22.4	613.5	80.2
Dec	13.0	36.9	22.8	613.9	78.1
1985 Jan	12.8	37.0	23.3	638.6	72.7
Feb	12.2	37.0	24.4	724.6	65.9
Mar	13.0	36.9	24.5	699.4	67.4
Apr	12.8	36.7	24.1	660.0	76.5
May	12.2	36.7	23.7	660.0	76.1
June	11.7	36.8	21.7	660.1	79.3
July	11.5	36.4	21.2	660.0	75.3
Aug	10.9	36.8	21.3	660.2	75.1
Sept			20.7	659.5	72.9
Oct					
Nov					
Dec					

67

There was an obvious problem in the barometric pressure readings before/after January 1985.

It should be recognized that January 1985 was the time of the changeover of field supervision. It appears that there may have been a change in the protocols at that time. This should be confirmed.

The Kenya project also has available one or more data sets of measurements on staff repeated over the life of the project. Obviously these should be examined to check for drifts in each of the parameters recorded. There may also be a data set incorporating measures of BMR in the same individuals in Kenya and in Los Angeles. If so, these should be scrutinized to see if there is a consistent difference.

Personal Conclusion: Until we can find the source of the unusual values in the Kenya data set, we are compelled to assume the worst--that the data are uninterpretable. ME must, as a matter of priority, examine the Egypt data to see if there is any possibility of a machine problem in that Beckman.

Other Matters

Dr. Neumann urged that ME provide additional guidance on the proposed structure of the methodology manuals--perhaps an annotated outline.

APPENDIX IV

Dr. Edward Stanek's Minutes of Data Analysis Group Meeting
Purdue University (6/25-6/26/86)

6/30/86 mex3.txt on MEX1

Minutes of Meeting of Statisticians 6/24/86 Purdue University

Present: Afifi, Beaton, Feinberg, G. McCabe, L. McCabe, Selvin, and Stanek.

There was initial dicussion over the meeting adgenda. Topics for discussion include questions on

1. The status of data
2. Construction of variables
3. Analytical strategies for each of food intake, morbidity, anthropometry, and psychology data. Other items for discussion include:

a. Response to Srinivasan's letter
b. Response to Dor^is Calloway's memo
c. Possible additional statisticians meeting prior to Nov 10-14 meeting.

A time table for sending results to EEP prior to the November meeting was made after consulting Hugh Horan. The time table is : Working copy of results to EEP

Target Date: 10/1/86
Last Acceptable Date: 10/15/86

Donovan Thompson and Keusch (Morbidity) will come from EEP. Srinivasan is also welcome to come, and the statisticians think he should be invited. Something more than potential for results should be presented at the November meeting. It was noted that the pressure for analysis should come from PI's, but in practice, has come more from the statisticians. If more resources become available (mabe 150K), Hugh H. reported via George B. that Dor^is C. agrees in principle to distribute the moneys within projects (presumably to be channelled to statisticans).

There was discussion by some of the projects on the status of diet and morbidity data.

Diet:

Egypt: Data is currently available for Dec 83-June 85, and Aug. 85. There is a problem with the Ramadan period, in which there is little data. Inquiries have been sent when Kcals differed by more than one standard deviation above the overall mean, to verify true values. This is a labor intensive process. There are other problems. About half the toddlers are not weaned at 18 months. G. McCabe will use some adjustment to diet data for this period, similar to that presented by Selvin at the October 85 meetings. There are other problems with lead males, since data from only one meal may be recorded with the other meals missing. Should this be "corrected"? To identify possible "outliers", Kcal measures larger than 2 x overall median were noted. The value of 2 x median and proportion of records with kcal measures larger than this are given below:

70

Target Type	2 x overall median	Percent of Subjects
Lead Males	4780 Kcal	2.0%
Lead Females	4426	1.7%
Schoolers	3498	4.3%
Toddelers	2108	6.5%

Potentially, such extremely large Kcal measures could be excluded, but more stringent criteria, excluding less cases may be desirable. In Egypt, there is a strong day of the week effect with higher intake on Friday. There is also a disproportionately larger number of Friday collection days among the high toddler data, suggesting that these data are real. The possibility of Windsorizing the extremely large values (replacing them by a maximum value) was discussed. As a result, the following criteria was established for cleaning Kcal data to be adopted by all projects:

Criteria for Range checks and Data Cleaning for Kcal

On an individual basis, calculate an upper limit equal to 2.5 x median Kcal for the individual. Also calculate an upper limit equal to 2.5 the overall median Kcal value for a target type. If a given Kcal measure for the individual exceeds either of these limits, then replace the Kcal measure as follows:

if Kcal > Max (2.5 x median Kcal over all target subects,
2.5 x median Kcal for the given subject)
Then Kcal = max (2.5 x median Kcal over all target subjects,
2.5 x median Kcal for the given subject)

Windows for Kcals for Toddlers

There was also some discussion for time periods in which data will be included. Windows of 14 days prior to 18 months and 14 days after 30 months were agreed upon for toddlers. The medians will be calculated for data in this period. All projects agreed to use these definitions for the protocol period.

The discussion naturally lead to extablishing standard methods for identifying outliers for other data. Decisions for other data and targets are as follows:

Criteria for Outliers for Toddler Weight and Height

Weight: Form a simple linear regression for each subject, and set the maximum (and minimum) weight as + or - 2 Kg from the expected regression line. If the measure exceeds these bounds, use the bound (minimum or maximum value) as the Windsorized value.

Height: Form a simple linear regression for each subject, and set the maximum (and minimum) height as + 3 cm or - 2 cm from the expected regression line. If the measure exceeds these

71

bounds, use the bound (minimum or maximum value) as the Windsorized value.

All projects should note whether or not there was a change in the way height was measured (ie. prone to upright).

Criteria for Outliers for Other Target Types

For other targets (schoolers, Lead Males, Lead Females, Infants), maximum and minimum values will be set as 3 x pooled within subject standard deviation from an individual subjects mean. If a value is outside this range, the value will be replaced by the maximum (or minimum) value corresponding to that subject. (Note that these criteria may be modified for infants, since such criteria were not established explicitly.)

Definition of Month

All projects will use 30 day months (hinged on the toddlers 18 month (547 day) birth date). Although not explicitly discussed, the criteria for the first and last month may be different so as to include the window. Thus, the first and last month would each represent 44 days of observation.

Each project will report the proportion of measures below, and the proportion of measures above these minimum and maximum values by variable and target type.

Kenya: They have all the data from the field. The data for final analysis should be ready by September.

Mexico: They have about 80% of all data, and have established a schedule for the remaining data. Nearly all data should be at U. Conn by September.

Animal protein data was discussed, since it was mentionned by Dor^is C.. Does each project have it? Is it worth calculating it? Suzanne M. will check on this and some decisions will be made. Each project will be informed of her findings.

Morbidity:

Afifi distributed a docuemnt on morbidity in Kenya that defines a morbidity index. The index attempts to create a scale for the degree of illness, using categories of chronic low grade illness, and significant illness. Charlotte N. feels the index may predict growth and the rating for illness may also be useful energy. (See attachment).

72

The morbidity variables from the February meeting are as follows:

	1.	Acute episodes
focus>	2.	Days sick
	3.	Duration per episode
focus>	4.	Days with Fever
focus>	5.	Days with diarrhea
	6.	Weeks with disease on household basis
focus>	7.	Morbidity index (as per Afifi)

There was some discussion of other probems in morbidity. Norge has derived four situations that would allow for different treatments of overlapping illnesses. For example, ther are progressive diseases, independent diseases, redundant diseases, etc.. There was much interest in progressive diseases. George M. distributed a copy of a correlation matrix for these morbidity variables and Kcal/kg for Egypts data (see attachment). It was agreed to focus on several of these indices, namely 2, 4, 5, and 7. Differences exist in Mexico's morbidity, since only illnesses present at the day of interview are included, and a list of symptoms was constructed which does not readily match the other projects.

Discussion of Possible Modeling strategy

Steve Selvin suggested using an autoregressive modeling strategy proposed by Rosner, Munoz, Tager, Speizer and Weiss, "The use of an autoregressive model for the analysis of longitudianl data in epidemiologic studies", 1985, Statistics in Medicine, Vol. 4. 457-467. There was enthusiasm among the statisticians for its use and subsequent discussion was aimed at defining a setting where it could be used. The model is a regression model and can be fit using standard package programs. The model discussed is as as follows:

Weight for subject i at time t =

Constant	+
Weight for subject i at time (t-1)	+
Kcals for subject i at time (t-1)	+
Morbidity indix for subj i at (t-1)	+
Baseline Kcals for subject i	+
SES for subject i	+
Sq root of household size for subject i	+
Baseline Morbidity index for subject i	+
error	

The discussion defined intervals for the times t-1 and baseline in which the variables could be constructed. A summary of the models proposed is included in a separate attached document.

72

Steve F. noted that it would be possible to separate subjects into groups by assuming the constant is distributed with parameters (A, B). Nick Joule has done some work on this. Also, Steve F. noted that this model is similar in some respects to models proposed by Laird and Ware.

Modifications to the model by Rosner et al. were introduced to account for different time intervals between consecutive weight measures. The model proposed by Rosner assumed equal time intervals.

The question arose as to whether food intake was measured when a subject was ill. All projects should establish whether the presence of illness affected whether or not food data was collected.

The regression analyses should be sex specific.

When three consecutive days of diet data are available, Steve F. argued that it is desirable to throw out the third day, so as to keep the same level of measurement error as when only two consecutive days of diet data are available. If only one day of diet data is present, it will be used. It was agreed to throw out the third diet data point, if such points exist.

74

Minutes of Meeting of Statisticians 6/25/86 Purdue University

Present: Afifi, Beaton, Feinberg, G. McCabe, L. McCabe, Selvin, and Stanek, and later Ted Wachs.

Letter from Srinivasan: George B. drafted a possible response. George M. will call Srinivasan and write a brief letter in response to the letter to Payne, commenting on the current status of the project. George M. will circulate the letter to other statisticians before sending it. Steve F. commented that the statisticians should think of Srinivasan as a colleague, and suggest that he could be invited to the November meeting.

Additional Stat Meeting: Dates for a tentative stat meeting were set as September 11-13, perhaps in LA. The purpose of the meeting will be to share the results and get the November report to EEP organized.

There was some discussion of the amount of effort required in data management to construct models like those discussed on 6/24/86. Ed S. proposed similar models based on constructing files using a 1 month data structure, as opposed to tayloring the models and time periods directly to the dependent variables (see last page of attachment on models). This would make fitting such models easier, but may make them less sensitive. Afifi said he will discuss the two approched with Ben B., and get an estimate of the difference in effort for the two approaches. He will decide on the approach he will take then. Ed S. said he will first use the 1 month structured approach, and use the alternative approach for weight if time permits. George M. will try both approaches on weight .

Models for psychology similar to the model for weight were proposed.

For Baily Psychology data, two possible models will be developed, one for Baily scores at 24 months, and one for Baily scores at 30 months. The 24 month model will use the 18 month measure as the lag variable, while the 30 month model will use the 24 month measure as the lag variable. The model for the 24 month Baily is as follows:

$$Y_{i1} = \alpha + \gamma_0 Y_{i0} + \gamma_1\left(\frac{Y_{i0}}{\Delta t_{i0}}\right) + \beta_{01} X_{i1} + \beta_{11}\left(\frac{X_{i1}}{\Delta t_{i1}}\right)$$

$$+ \beta_1^* Z_{i1} + \beta_2^* Z_{i2} + \beta_3^* Z_{i3} - \beta_4^* Z_{i4} + e_{i1}$$

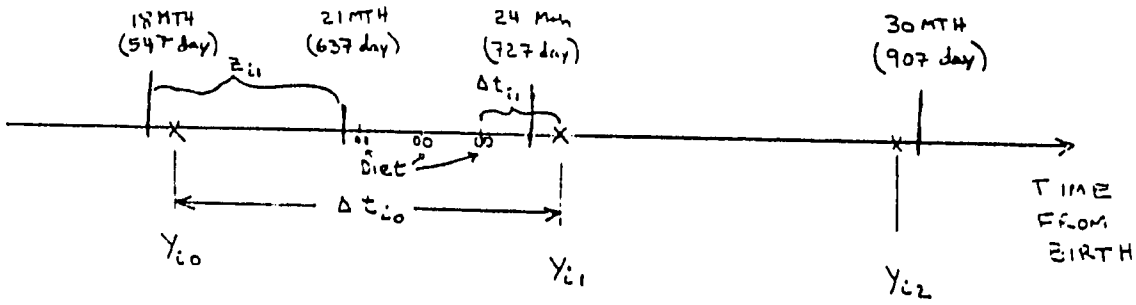
75

where Y_{i1} = Bailey for toddler i at time 24 months

Y_{i0} = Bailey for toddler i at time 18 months

X_{i1} = Most recent 2 day diet data average preceding 24 mth psyc measure

Z_{i1} = Average of 18-21 months intake data



and where Z_{i2} = SES

Z_{i3} = Morbidity Index

Z_{i4} = Sq root of household size

A similar model will be constructed for Bailey scores at 30 months.

Psychology

Ted Wachs joined the group from 10:00-12:00 AM.

The Bailey scores used by Egypt and Kenya are raw scores, not normalized to US standards. The raw scores should increase over time. The Bailey scores are conceptually comparable across projects. The psychologists are more interested in how the environment affects behaviour and development than the relationship between nutrition and behaviour. Care giving responsibility and the interaction of nutrition and care giving are variables of interest. There are two or three variables that can be constructed by the psychologists from care giving that would be hypothesized to be predictive of the Bailey.

76

Data collected in the protocol was reviewed.

18 months	24 months	30 months
Toy Play Interaction	Toy Play Interaction	Toy Play Interaction
Styles of Behaviour	Styles of Behaviour	Styles of Behaviour
Bailey Mental	Bailey Mental	Bailey Mental
		Bailey Motor
Memory	Memory	
Persistence	Persistence	
Foresight	Foresight	

There was some discussion of possilbe relationships or hypotheses that Ted thought would be of interest. The most sensitive outcomes are memory, persistence, foresight, and toy play scores. A change down in toy play score represents an improvement in score. A potential hypothesis is "Do the same nutritional influences predict the mental and motor Bailey scores?". Morbidity may be a predictor of Bailey score. A measure of chronic morbidity may also be of interest. Morbidity variables that may be related to psychological measures are fever, and gastrointestinal illnesses, as opposed to upper respiratory illness. Vision and hearing problems may have been noted somewhere and toddlers with such problems should be treated separately. Parent IQ should be factored in the models. Parent IQ may not be reflected in toddlers Bailey scores, but may be related to toy play. Ted estimated that if psychologists had clean data for about a month, three key variables could be identified from the toy play data.

The CRSP question is : What is the role of chronic mild malnutrition on cognative development?

Important covariables for Bailey scores are:

- Morbidity of primary care giver (mother or older sib)
- Childs own morbidity (preceeding month and chronic)
- Parents cognative status (Global IQ)
- Food intake for primary care giver (preceeding month and overall)
- Food intake for child (preceeding month and overall)
- SES

77

Height and weight may be of less interest. The major independent variable are the care giving variables. Across project comparisons are of interest in seeing whether predictors are the same in the three settings. There was agreement that for the November meetings, psychology analyses will not be planned beyond the toddler data.

Data on newborns (Braziltons) was mentioned, but not discussed.

Models for other Dependent Variables

Other models were defined using the autoregressive model framework. The first step in fitting these models is to construct a data base on a monthly basis. This data base should include the average measure in the month, the average time at which the measure was made in the month, and the number of measures in the month. Variables included in these models are as follows:

Variables

WT = Weight
HT = Height
FI = Food Index, either Kcal or Kcal/Kg
MI = Morbidity Index, or other morbidity variable
HHS = Square root of household size
MI_B = Baseline (3 month) morbidity index
FI_B = Baseline (3 month) food index
HT_B = Baseline (3 month) height index
%Fever = Percent of days in month with fever.

The models to be fit are specified by listing the dependent variable, first order auto regressive independent variables, and other static independent variables. Models to be fit include the following:

Dependent Y	Auto-Regressive X	Static Independent Variable Z
WT	FI, MI	SES, HHS, MI_B, FI_B, HT_B, WT_B(?)
HT	FI, MI	SES, HHS, MI_B, FI_B, WT_B, HT_B(?)
MI	FI, WT, HT	SES, HHS, MI_B(?), FI_B, HT_B, WT_B
%Fever	FI, WT, HT	SES, HHS, FI_B, HT_B, WT_B, %Fever_B(?)
FI	MI, WT, HT	SES, HHS, MI_B, FI_B(?), HT_B, WT_B

There was some discussion of basic descriptive statistics to be presented by all projects. Ed S. suggested all projects compile the number of subjects enrolled in the each project, and the number of subjects used in the analysis. George B. noted that subjects are linked according to the protocols. For example,

Toddlers --> Lead Female, Lead Male
Schooler --> Lead Female, Lead Male
Pregnant --> Lead Male, Lead Female, Infant.

Further discussion indicated that such tables would not be easily constructed (since they would have to be data type specific), and may not be that useful. They will not be constructed.

Discussion of Doris C. Memo

There was discussion of Doris Calloway's May 1 memo on analysis ideas. Some of the ideas suggested in the memo overlap the models that have been proposed in this meeting. Other analyses that are suggested are iterative in nature, requiring constant contact with the investigator. This isn't possible (except within individual projects) within the framework of the CRSP with Doris. Steve S. will discuss this with Doris.

That's it.

Ed Stanek

79

APPENDIX V

Documents from Data Analysis Group Meeting
Chicago (9/12/86)

a. Dr. George Beaton's Meeting Notes

MEETING NOTES

George H. Beaton

Meeting of the Data Analysis Group

Chicago

September 12, 1986

81

The meeting was held in the Traveloge Inn at the Chicago airport. Present were George and Linda McCabe, Ed Stanek, Steve Feinberg, Gail Harrison and George Beaton. The McCabes, Gail Harrison and George Beaton also met the evening before. No one was present to represent the Kenya project however there was a prearranged telephone conference between George McCabe and Afifi followed by a short discussion between myself, Afifi and Ann Coulson on other CRSP matters.

Minutes of the meeting will be prepared by Ed Stanek and will be sent to George McCabe for review and distribution. What follows is a short summary of some of the major areas of discussion, not necessarily presented in the order of discussion nor in any particular order of importance.

Integration of Psychology/Cognition into the Interproject Analysis

It appears that at this time the only project reporting major advances in the intraproject analysis of psychology data (for toddlers) is Kenya. It is understood that details of this work have not yet been shared across projects, beyond the verbal report presented at the last SCB meeting and hence can make little contribution to the planning of analyses by other projects. The Egypt project does not yet have all of the data available for analysis although preliminary analytical work is underway. The Mexico project is planning analyses of some aspects of the psycholody data.

There was a strong feeling that an interproject approach must be started as soon as possible. Two recommendations toward this purpose were offered:

1. It was recommended that the ME take action to try to initiate a meeting of representatives of the three projects to discuss their analyses of toddler psychology data. The meeting should include also at least one PI who is not herself presenting psychology analyses and at least one of the project statisticians.

   The objectives of the meeting should be: i) to ascertain the specific analytical interests and approaches of the individual projects, and to share experiences in the interpretation of data, ii) to begin the specific development of both explicit question definition and analytical approach for parallel interproject analysis of the central CRSP questions relating to psychology.

2. It was recommended that the November meeting include a day when project psychologists, PI's and statisticians could meet to specifically deal with interproject analytical approaches to psychology analyses. This should be prior to the arrival of EEP members and should focus on the interproject rather than intraproject analyses. The outcome of the meeting would then be reported in the SCB meeting with the EEP representatives as well as guiding analyses in subsequent months.

Experience in Application of the Autoregression Model

Much of the day was spent in reviewing and discussing experience of the Egypt and Mexico projects in applying the autoregression model to toddler data. The data sets underlying these analyses remained incomplete (i.e. further updates in the data and hence in the analyses, were to be expected).

Projects had abandoned the interim decision to Winsorize data and instead had adopted rules for rejecting data (declaring it as missing). The numbers of data that were being rejected under these new rules were indeed quite small. Attempts to apply the rules of defining limits in terms of deviations from an individual's regression over time were found to be too sensitive to outlier effects (when the outlier was at the beginning or end of the series it exerted too strong an influence). The final rules adopted by each of the projects will be explicitly defined. The Mexico project had updated analytical files to take into account the new decision and had then updated files to take advantage of new food intake data but had then found that the identifiers on the new food intake data did not always match the old data. This is now being reviewed/corrected but it did set back the analyses (and implies that present analyses are not final).

A small effect of food intake on weight gain was discernable in the analyses. In discussion it was pointed out that the style of presentation of results would be important. The dominant predictor of body weight was, as would be expected, a cluster of measures of previous body size (weight, height, baseline weight and height). R squares with these varaibles only were in the order of 0.7 to 0.75. The additional predictive power of adding prior food intake increased the R square only marginally and hence appeared to be of no importance. However, if it is recognized that what we are attempting to examine is the influence of "deviations" of food intake on the "deviations" in weight after giving recognition to the expected predictive power of previous body size, then we are attempting to isolate the sources of the unexplained 25-30% of variance in weight and the contribution of the food intake variable becomes more impressive.

There was discussion of the possibility of using parameters from the individual regressions, including residuals, as variables in the autoregressive models. This was subsumed by a later discussion of the use of subject as a variable (see below).

There was discussion of the possible effects of colinearity in the models and examination of results to see if such effects were apparent in the analyses. It was not felt to be a problem.

The Egypt project had introduced dummy variables for missing data and by this technique had maintained a higher sample size for analyses. The Mexico project had selected and examined only the data sets that were "complete" with regard to the groupings of variables to be examined. For full variable sets, this would lead to a considerable attrition in sample size and hence loss of power. There was considerable discussion of the impact and interpretation of the dummy variables without a clear resolution but with a decision to run parallel analyses with and without dummy variables to really test their impact on the autoregression results.

All agreed that there was still considerable difficulty in understanding exactly what the model was doing and was portraying, particularly when many variables were included. It was agreed that any final presentation of the model should include as few variables as possible, but including those that, by prior testing, made significant contributions to predictive power or were clear tests of the CRSP hypotheses.

The general feeling was that the analytical model was still encouraging but that some additional refinements, based on the experiences with approaches tested, would be advantageous. Both the Egypt and Mexico projects had undertaken considerable testing of varaints on the model, always using the basic autoregressive structure. Some specific suggestions included:

- o Mexico might attempt the inclusion of dummy variables for missing data in parallel to the approach of Egypt
- o Egypt might run the regressions using only "complete" data sets and omitting dummy variables to see if this appreciably reduced the attribution of predictive power to the intake (and other) variables.
- o Recognizing that SES variables appeared to have little predictive power, but that in the Egypt data a function of household size did seem to have effect, it was suggested that both projects run the models without SES variables but with the Household Size variable included.
- o Inclusion of subject as a variable and determination of whether or not this captured the same variance as the cluster of body size descriptors was seen as desirable. The Mexico project would include height as well as weight in the baseline measures but could not include height in the prior month measures (data not collected consistently). The Mexico project was also examining the inclusion of "cohort" variables/running models independently by cohorts defined either in terms of entry date or by community in an attempt to establish whether there were any real cohort effects that might have to be taken into account.

- o Further development of morbidity variables for inclusion in the model seemed to be indicated, recognizing that these variables could not be parallel across projects because of important differences in the nature of the morbidity data collected.
- o A goal might be to isolate those variables that had significant predictive power and to then run, for presentation, these models. An accompanying narrative would stipulate the other variables tested and found to be non-contributory.

It was agreed that the Mexico and Egypt projects would work closely together in developing and refining models that could then be run by the Egypt project and then sent out by George McCabe to both the Mexico and Kenya projects with a request that they be run in as close to a parallel manner as possible. The final runs of these models would then be consolidated for presentation in November as an interproject analysis (recognizing that this might not be the final analysis to be performed in this area). [See below for expression of concerns about the logistics of doing this.]

It was suggested, and agreed, also that one of the projects should test the model using a longer interval of food intake data to predict weight (e.g. mean food intake across the prior three months). This was recognized as involving a trade-off of error terms and might or might not be advantageous. That is, the approach should provide a more reliable estimate of average food intake in the period, and hence reduce the impact of this error term on the regression coefficients while at the same time reducing sensitivity to more recent deviations in food intake - it would be asking questions about the relationship of more chronic food intake to subsequent body weight.

Morbidity as an Outcome

Preliminary runs suggested that the autoregression model, as applied to weight data, might not be the most effective model for morbidity data. Certain classes of morbidity did indeed predict subsequent morbidity, as would be hoped. However, a number of issues were seen in relation to the use of predefined time intervals (one month windows) for the definition of a morbity outcome variables. The classification scheme developed by the Kenya project did not seem to be applicable to either the Egypt of Mexico projects. In preliminary work, the Egypt project had devised a scheme for classifying individual episodes of illness into 4 or 5 categories including as one "progressive illnesses" and including also categories which differentiated between overlapping illnesses prodromal or multiple successive appearance of expected symptomilogy of the same syndrome.

The Egypt project plans to explore the use of these classifications in a model in which the disease episode would set the time intervals and the analytical question might relate to the differentiation of conditions that predicted not only the occurrence of a disease category but also whether the disease did or did not progress.

There was discussion also of the possibility of categorizing individuals across the whole year into categories of patterns of illness, e.g. realitively illness free, chronic illness, progressive illnesses, and to then analyse these in relation to other variables captured across the longitudinal data set. That is this would be an analysis which made use of the longitudinal information but applied a "cross sectional" approach.

Cross Sectional Descriptor of Morbidity

It was recognized that in comparing and contrasting the populations studied in the Nutrition CRSP, it would be desirable to generate at least one morbidity descriptor that could be applied in parallel across all projects, given that each project had elected to collect morbidity data in a different, non-compatable, manner. This would not preclude individual projects applying other descriptors but would yield at least the one common variable. A least common denominator was defined that could be developed by each project.

> It was recommended that each project report the proportion of toddlers (and later other target individuals?) who were reported to be ill on the day of the morbidity recall visit and that this be displayed as an interproject comparison of morbidity prevalences.

BMR/RMR Data: The Problem of High RQ's: The Nestle Foundation Studies

In the telephone conference with the Kenya project it was reported that the source of the problem in RMR data (the high RQ's) had been isolated in consultation with Beckman. It appears that indeed it was an equipment problem and more specifically an error in the computer programming. Apparently the algorithms did not take into account humidity in deriving gas volumes. The erroneous data can be detected in the machine printouts and manual corrections can be made. That is, the Kenya project and the Egypt project can correct the RMR data but it will involve major work and costs for reprocessing the extensive data sets.

> It was recommended that ME, or the individual projects, but preferably the former, seek legal advice and then seek full compensation for total costs of all projects from the Beckman company.

The costs to be recovered should include not only the costs of data rentry but also a settlement of costs related to the extensive examination of data to isolate the nature and source of the problem. In addition, equipment modifications for all three Beckman units purchased by the CRSP should be demanded. ME might wish to solicit the cooperation of US AID in its approach to Beckman.

Dr. Beaton advised projects of the reported results of the Nestle Foundation projects on energy balance in pregnancy and lactation, noting that a summary report had been distributed to projects. For

analysts he pointed out that there were a number of implications not the least of which was the apparent fact that food intake increased very little in pregnancy but more materially in lactation. BMR may fall in early pregnancy (and then progressively rise in later pregnancy - but perhaps with limited apparent change in relation to non-pregnant values) and to be significantly lower in lactating women (first few months anyway) than in non-pregnant, non-lactating women. This had implications for the need to categorize lead females in any presentation of food intake or BMR data as well as clear implications when we move to examination of the pregnancy-lactation-infancy hypotheses.

Planning and Organization for the November Meeting

A preliminary agenda proposal, prepared and distributed by Lindsay Allen as Chairman of the SCB, was brought to the meeting by Ed Stanek. There had been discussion of this proposal between Lindsay and Gail. The discussion in the present meeting led to identification of several concerns about the proposed agenda and about broader issues relating to the meeting.

Considerable concern was expressed about the planning of the November meeting and the related planning of analyses and data flow in preparation for that meeting. It was absolutely clear that there was a need for defined leadership in this - and that this had not been defined to date. The Data Analysis Group members noted that this was likely to the last time that the analysts and PI's would meet together within the life of the CRSP (except for a final meeting involving presentation of the final reports). The opportunity for discussion of analytical approaches must not be wasted and this would require a tight structure for the meeting as well as a carefull planning and assignment of responsibilities in preparation for the meeting. The meeting would serve also as a presentation of progress to EEP members but this was not the primary purpose. The EEP needs would be met most effectively by a meeting that addressed analysis issues in an ongoing mode - progress, issues and plans.

Among the issues that were apparent to the statistical group were matters such as:

- What descriptive data were to be assembled and presented to support analytical presentations (analyses performed and analyses planned)?
- Who was to assume responsibility for announcing the data format (for projects to output) and who would accept responsibility to receive and assemble these data?
- Who was to prepare the consolidated report on analyses performed and who would actually present this report?
- How would reporting (in meeting and in documents) be divided between the collaborative and project-specific analyses? It was clear that there would be parallel analyses run as a part of the collaborative inter-project analyses and there would also be complementary

intraproject analyses that offered specific variants on the general models.

- Who would take responsibility/be given responsibility to prepare interpretations of the analyses for presentation and discussion in November?

The point was that there was urgent need for logistical planning for this meeting and for definition of authority/responsibility if the meeting were to be productive in the relatively short time available. In this regard, it was felt that the last meeting, in Los Angeles, had been less productive than it should have been in terms of real progress on analysis issues.

> It was recommended, as a matter of urgency, that there be a telephone conference or other communication between the three principals in the forthcoming meeting,
>
> Hugh Horan - on behalf of the Management Entity
> Lindsay Allen - on behalf of the SCB
> George McCabe - on behalf of the data analysts
>
> with the objective of defining, very explicitly the objectives of the meeting, the format of the meeting, and the logistical arrangements for preparation for the meeting, including the assignment of overall and specific responsibilities and authorities.

There was discussion of the benefit to projects of assigning to Berkelely a shared responsibility in preparation of first level analyses of data or other related preparatory/background information. Specifically the problems and loads imposed in obtaining common data outputs from individual projects at the same time that projects were attempting to complete analyses was seen as an issue that the CRSP should be competent to avoid by this stage of the program. That was not an implication that comparative descriptive analyses should not be prepared but rather than there must be more cost-effective ways of doing it. The analysts present were not in a position to offer recommendations on this matter, recognizing that it was an issue on which the SCB had expressed strong feelings. Nevertheless one member of the group (not GHB) expressed the view that the projects were greatly underutilizing the human and other resources offered by Berkeley - and that this was probably operating to the detriment of the whole CRSP.

Future Involvement of Kenya Project in Analysis

In the telephone conference call, the statisticians were made aware of the possibility that the Kenya project might have to withdraw from further analyses (and hence we could not be assured that parallel analyses would be performed for incorporation in a report for the November meeting). This was seen with regret and concern but was seen also as something to be addressed between Management and the Kenya project. It was suggested that any documents prepared to support the presentation of data and collaborative analyses should include spaces for the Kenya data whether or not the data were available at the time of preparation.

ANNEX TO BERKELEY

Attached to this report are copies of documents tabled at the meeting. These include the tentative agenda for the November meeting sent by Lindsay Allen, and analytical results presented by George McCabe and by Ed Stanek.

PURDUE UNIVERSITY inter office memorandum

To Afifi, Beaton, Fienberg, Harrison, Horan, Selvin, Stanek

From George McCabe

Date September 10, 1986

Subject Chicago Meeting

Beaton, Fienberg, Harrison, Stanek, G. McCabe and L. McCabe will meet at the Travelodge Motel near the O'Hare airport on Friday, September 12 from 9 a.m. to 6 p.m. The following is a first draft of an agenda:

1. Results of analyses proposed at the June meeting including identification of bad data and autoregression models.

2. Other analyses in progress which will be ready for the November meeting.

3. Outline of formats for presentation of summary statistics at November meeting.

4. Psychology analyses.

5. Comments on RMR/BMR methodology (Beaton).

GM:tc

From L. Allen.

mexico
nov1.txt on \mexico\minutes disk # 1
9/11/86

Plan for November Meeting

Monday Nov 10. PI's and Stats meet to discuss adgenda

PI's meet Discuss analyses
Stats meet Discuss analyses

evening: Psychologists arrive Ted Wax, Marian Sigman, Tiffany Fields

Tuesday Nov 11. Presentations:

Stats: Parallel reg models for toddlers (no psych data)
PI's: Several analyses that have been conducted
At least two analyses conducted in parallel at two projects. (Adolfo and Nel on Morb)

2 EPP people present: Jerry Kursh
Donavon Thompson

Concerns:
1. Phrase hypotheses with CRSP language (marginal energy intake (indep) on function)
2. Show we are doing analyses in common
3. Illustrate how we constructed parallel morbidity variables.
(Each project get # days sick on days visited)

Wednesday Nov 12. PIs, Stat, EPP

Cognative Analyses presented (by PI's)

Afternoon: Psychologists off to come up with consensus for remaining psych analyses
Include George Beaton, Hugh, mabe Gretal

Open time, Stats, PIs

Thursday Nov 13. Psych people present united front
Stat meet
PI meet

Friday Nov 14.
Overall Summary
Stats leave

APPENDIX V

Documents Relating to Data Analysis Group Meeting
Chicago (9/12/86)

b. Dr. Gail Harrison's Trip Report with Appended Articles

TRIP REPORT
Nutrition CRSP
Chicago, September 12, 1986

Gail G. Harrison
University of Arizona

At the request of Management Entity, I traveled to Chicago to represent the PI's at a one-day Statisticians' Group meeting on September 12th. Others present were George McCabe, Ed Stanek, Linda McCabe, Steve Feinberg, and George Beaton. Dr. Stanek took minutes, and Dr. Beaton represented ME, so these notes are primarily for the purpose of communicating with the other PI's. Unfortunately, no one from the Kenya project was able to attend, although a brief telephone conversation with Dr. Afifi at UCLA had been pre-arranged and occurred during the day.

The meeting was, in my opinion, most productive and timely. As always when in company such as that named above, I came away with my mind stretched and some new viewpoints on the data.

Major topics discussed included the following:

1. Detection of errors and "cleaning" of anthropometric data.
The Egypt Project statisticians reported that the method agreed upon in theory at the June meeting (refer to Stanek's minutes of that meeting) had proved less than entirely satisfactory, both in terms of identifying outliers and in terms of identifying points in error which were not true outliers. We discussed a paper which I had brought (Duquet et al., enclosed) on detection of errors in growth studies, which discusses a technique we have long used, of scanning the data not only for outliers but for unusual relationships between variables which should be highly correlated.

2. Discussion of progress in applying the autoregression models of Rosner et al. (see enclosed paper) to the toddler anthropometric, morbidity and food intake data, as agreed upon in the June statisticians' meeting. Both the Egypt and the Mexico projects have done some initial analyses, and these were presented and discussed. The approaches have been somewhat different in two respects. One is the variables tried in the model in addition to those which had been agreed upon. The other major difference was that the Mexico group had run the model essentially on only complete cases, while the Egypt group had included dummy variables to allow honest inclusion of cases with incomplete data. It was agreed that both groups will do some further runs, using each others' approaches, to produce more directly comparable material. These will be incorporated, along with whatever UCLA is able to provide, along with some background and explanation of the model which Dr. McCabe will draft, in the statisticians' report to be submitted to ME by October 15 for forwarding to the EEP.

We discussed a number of substantive issues about the models and their application to the data. I think I was able to provide some perspective on the following items:

-- the lag time used in these initial analyses is probably too short to see the effects we are looking for. Certainly this is true for height (where, in my opinion, the model should be predicting the slope of the line of height change over the year or a least over 6 months, rather than height in one-month intervals). It may also be true for weight.

-- the baseline morbidity variable is measuring (at least in the Egypt data) only season since morbidity is highly seasonal and the baseline variable (defined in the June meeting as sick days in the first three months in the study) will be closely related to the month in which the toddler entered the study. In further discussions with George and Linda McCabe and Nell Kirksey on 9/13, we decided to try instead a variable which expresses the toddler's baseline morbidity (1st three months) in relation to that of other toddlers in the same three-month period.

The idea of "progressive" or "progressively severe" morbidity may be possible to address in the Egypt dataset by examining overlapping diagnoses (i.e., overlapping in time). I will pursue classification of these events with Dr. Sieber and several independent clinicians. Especially with regard to respiratory illnesses, this may be an especially useful category of illness experience. The Mexico project is not able to extract comparable data; we are not sure whether the Kenya project could, but based on my last conversations with Charlotte I don't think so.

--The different nature of the morbidity data bases in the three projects. It was agreed that the Egypt project and Mexico projects would run a parallel analysis of prevalence of various illness, i.e., ill on the day of the interview.

Both projects found, not surprisingly, that about 70 percent of the variance in weight at a given point is predicted by weight one month previously. The ability of other variables (morbidity, food intake etc) to explain the remaining variability is the appropriate focus. At least in the Egypt project analyses, it looks as if biologically appropriate if small effects may be present, both for some types of morbidity and for food intake. I am encouraged.

* PI's: READ THE ROSNER PAPER.

3. Psychology analyses were discussed at some length, with regard to how best facilitate rapid progress in this area. The group decided to recommend to ME that the group of three project psychologists be convened at the earliest possible date, with a PI and a statistician present (possible at Connecticut, with Stanek?) to discuss their approaches to the data and agree on an approach. The group also felt, and I agree, that it will be important to have a day for the psychologists, statisticians and PI's to discuss the analysis of these data together at the November meeting in Berkeley -- before the EEP arrive. Dr. Beaton was to convey these recommendations to ME.

The nature of the statisticians' report to EEP vs. the information to be presented to be presented in projects' annual reports was discussed. The general feeling was that project annual reports should contain as much descriptive data as possible, while the statisticians' report should be pretty much strictly analytical in focus, and not too long.

There was some discussion of the RMR data, spurred by the conversation with UCLA folks in which they indicated that they have identified some problem internal to the Beckman's programs which may affect interpretation of data. We tabled the discussion pending more information.

George McCabe, Lindsay Allen and Hugh Horan need to talk about finalizing the November agenda, who runs the meeting, etc. George McCabe will get in touch with Lindsay Allen and Hugh Horan.

ER&P:gght0919.crs

hermaphrodites. *American*

mouse. *American Journal of*

7, 315–321.
mongol imbeciles. *Annals of*

(palmar) interdigital area
Research, 11, 36–42.
dermatoglyphics. *Journal of*

23, 49–69.
dermatoglyphic asymmetry.

the testicular feminization
ndon, B, 259, 187–204.
Neurology, 13, 235–238.
order. In *Developments in*
Vickers Lectures of the
London. Sydney, Auckland:

word: Blackwell Scientific

n testicular feminization.

phic findings of palm and
rdstm, 17, 729–740.
ation testiculaire totale et

childhood and Adolescence

Guy's Hospital Medical

stikulären Feminisations-
ive als auch qualitative
nds, die die vorsichtige
eit der Einfluß des
n Schleifenmuster und
einflußt werden mag; im
er bis zu einem gewissen

des doigts et des paumes
aire, tirées de 28 lignées
des conclusions tentées
ies sexuels est impliquée.
peuvent être influencés
les crêtes digitales et.
sure par les androgènes

ANNALS OF HUMAN BIOLOGY, 1979, VOL. 6, NO. 5, 431–441

A method for detecting errors in data of growth studies

W. DUQUET, F. DE MEULENAERE and J. BORMS
Laboratory of Human Biometry and Movement Analysis, Vrije Universiteit Brussel, Hilok

Received 11 July 1977; revised 6 February 1979

Summary. The paper describes a new method used for data cleaning in growth and development studies. The method is essentially based on the assumption that an extreme value of a certain variable might be suspicious if other highly correlated variables show low compatibility with it and that, on the other hand, an extreme but possible value tends not to be erroneous if it is reinforced by another extreme but possible value of a highly correlated variable. Each variable was therefore included in at least two ratios with the highest correlated variables. A program was designed to detect extreme values of individual variables and extreme ratio values. This procedure with ratios was used to help to detect possible errors and discriminate them from true extreme values. Seen in the light of corrected data files against existing data files, the number of corrections was approximately 4·4%. If the real number of corrected errors is compared to the total number of subjects, this percentage reached 5·5%. If, when correcting, the real value was not detected with certainty, the erroneous value was then simply deleted. From our results, there is reason to believe that with this method few detectable errors will escape from the cleaning procedure. Conclusions for future correction procedures and for future growth studies in general, are also given in the paper.

1. Introduction

Population studies inevitably lead to different kinds of data error. Detection and correction of these errors is important for several reasons.

(*a*) If a correction procedure is possible, it should be done unconditionally in order to violate scientific integrity.

(*b*) Such procedures as removing extreme values are insufficient for eliminating all types of error. For example, a variable could show a not too extreme value for a particular person, but this value should be rejected if other highly correlated variables are incompatible. On the other hand, an extreme but possible value should be retained if it is reinforced by another extreme but possible value of a highly correlated variable.

(*c*) Data errors may seem unimportant when comparing their relatively small number to the complete data mass, but they can strongly influence the analysis of small groups of extreme, but true values which are of great importance.

(*d*) Correcting data errors saves valuable programming and computing time in further analyses.

The search for possible data errors is difficult and careful attention and caution is necessary. Basically every piece of collected information could be wrong, with errors occurring anywhere from original recording sheets to data tapes or discs.

Since there is little literature available on error detection, we designed our own system and applied it to the data of the Belgian "Performance and Talent" project (Hebbelinck and Cliquet 1970), which is cross-sectional in nature. However, the following method is also applicable to longitudinal designs. A survey of the procedure follows.

0031–4460/79 0605 0431 $02.00 © 1979 Taylor & Francis Ltd

96

2. Basic conditions

The procedures described below proved to be applicable for error detection and correction in a data collection with the following five characteristics.

(*a*) A large number of subjects (in our case about 7000) was involved.

(*b*) The information was written on individual data sheets at the time of measurement. Later it was recorded on magnetic tape.

(*c*) Each subject occupied several formated lines on the tape.

(*d*) Each line of the tape contained the subject's identification number and a line number.

(*e*) Other sources of information regarding the subjects were gathered in some cases: somatotype photographs, left hand X-ray, finger and footprints. These sources served as support data (from now on called 'witnesses'), in addition to other information, such as the original data sheets, frequency distribution tables and percentile scales. However, it was kept in mind that this information too could be erroneous.

3. Description of the work scheme

Preparation of information sources

First, all available 'witnesses' were classified to permit easy access to this information.

Execution of the program for detecting classification errors

A test to detect errors against the classification principles of the data system was carried out. A computer program was employed to ensure that no double use of identification number occurred, that each subject occupied the normal number of data lines, and that each subject's data lines were placed in the right order.

Corrections of classification errors

The errors detected above were immediately corrected on the tape, in the following sequence:

(*a*) selecting the subject's original data sheet,

(*b*) erasing from the data tape the complete set of information of each subject concerned,

(*c*) auding the information of the original sheets to the data tape. In some cases, a new identification number had to be given before adding the original information to the data tape.

Execution of the program for detecting extreme data values

A program (see figure 1) was written to detect extreme values for all variables. The program calculated means, standard deviations and tolerance levels for all variables. It was clear from the beginning that not all variables could be treated the same way, due to the specific form (coded or not) and the specific range of each variable. For this reason and for programming convenience, it was decided to use the same small limits ($2{\cdot}575\sigma$) for each variable, i.e. to obtain a large number of extreme values ($2{\cdot}575\sigma$ corresponds to a 2-sided $P<0{\cdot}01$ in a normal distribution).

A simple run of the single variables, indicating extreme values, was not enough. An erroneous data value may be thought to fall within the tolerance interval of the variable considered, yet may not be acceptable when compared to the same subject's values for other related variables. For this reason, the procedure described for detecting extreme

values was followed for ratios. Each variable was included in at least two ratios with the most highly correlated variables, and extreme ratio values were listed. The program output displayed the following items (see table 1):

(*a*) the subject's identification number,
(*b*) a number identifying the variable(s) or ratio(s) concerned,
(*c*) the value(s) exceeding the limits,
(*d*) the group mean for this variable or ratio (including erroneous values),
(*e*) the percentage deviation from the group mean of the subject's value.

First screening: deletion of non-errors

In a first screening, all extreme values that could not be classified as certain or possible errors were eliminated from the listing. While deciding upon the degree of eccentricity of a certain value, the specificity of the range of each variable or ratio was respected by using frequency distributions or percentile scales where possible.

Important criteria to decide whether a value should not be considered as an error were:

(*a*) if in the same subject, more than one variable showed extreme values, and if a logical commonalty was found in the nature of these variables *and* in the direction of the eccentricities;

(*b*) if an extreme value for a certain ratio was not accompanied by an extreme value of one of the single variables used in the ratio, nor by another extreme ratio value where one of the same variables was involved.

If any doubt occurred, the subject remained on the list of possible errors, and the individual scoring sheet was consulted.

Second screening: control with original data sheets

In a second screening, all remaining values were compared to the subject's original data sheet.

(*a*) If the comparison revealed punching errors, this was noted on the data sheet, and the subject's number was eliminated from the listing of extreme values.

(*b*) If the comparison revealed clearly correctable errors on the data sheet, these errors were immediately corrected on the sheet, and the subject's number was eliminated from the listing of extreme values.

(*c*) If the comparison revealed possible or certain errors that could not be corrected immediately, and the right witness was available, then the witness was requested, the subject number was kept on the list, and the data sheet was not altered.

(*d*) If the comparison revealed possible or certain errors that could not be corrected immediately, and the appropriate witness was not available, then the following situations could occur:

1. If a single variable showed an extreme value not accompanied by an extreme value of a related ratio, the single extreme value was considered as an error only if it greatly exceeded the limits of the variable's frequency distribution.
2. If an extreme value of a single variable was accompanied by extreme values of related ratios, and the extreme values reinforced each other, the variable in question was almost always considered erroneous.
3. If an extreme value of a certain ratio was accompanied by an extreme value of another related ratio, the common variable was almost always considered erroneous.

Figure 1. Computer program "DACOR"

```
      SUBROUTINE DACOR(TOL,TOLR,IR,IV,IP)
      COMMON BL RM(100,100),RMED(4,100),ID(100),IRA(2,50)

C  SUBROUTINE DACOR (THIS ROUTINE IS WRITTEN BY IR. F. DE MEULENAERE,
C                    LABORATORY OF HUMAN BIOMETRY AND MOVEMENT
C                    ANALYSIS. VRIJE UNIVERSITEIT BRUSSEL)

C  A) DIFFERENT PARAMETERS
C  ----------------------
C     1) TOL  =TOLERANCE LEVEL FOR ALL SINGLE VARIABLES.
C             IF 'XV' IS THE MEAN VALUE OF A CERTAIN VARIABLE, 'DV'
C             ITS STANDARD DEVIATION AND 'TOL' THE TOLERANCE LEVEL,
C             ALL VALUES AT THE OUTSIDE OF THE INTERVAL (XV+TOL*DV,
C             XV-TOL*DV) ARE CONSIDERED AS EXTREME VALUES AND
C             WILL BE PRINTED OUT.

C     2) TOLR =TOLERANCE LEVEL FOR ALL RATIOS.
C             IF 'XR' IS THE MEAN VALUE OF A CERTAIN RATIO, 'DR'
C             ITS STANDARD DEVIATION AND 'TOLR' THE TOLERANCE LEVEL,
C             ALL RATIOS AT THE OUTSIDE OF THE INTERVAL (XR-TOLR*
C             DR,XR-TOLR*DR) ARE CONSIDERED AS EXTREME RATIOS AND
C             WILL BE PRINTED OUT.

C     3) IR   =NUMBER OF SUBJECTS (NUMBER OF ROWS IN THE
C             MATRIX 'RM').

C     4) IV   =NUMBER OF VARIABLES.

C     5) IP   =NUMBER OF RATIOS.

C  B) MATRICES AND VECTORS SET BY THE USER(COMMON BLOCK 'BL')
C  ---------------------------------------------------------
C     1) MATRIX RM(100,100) :  EACH ROW CONTAINS ALL INFORMATION
C                              (VARIABLES+RATIOS) OF ONE SUBJECT
C                              (TOTAL=IR; MAX=100). EACH OF THE FIRST
C                              IV COLUMNS CONTAINS A VARIABLE; THE
C                              FOLLOWING IP COLUMNS CONTAIN THE
C                              THE DIFFERENT RATIOS (TOTAL=IV+IP;
C                              MAX=100).
C                              REMARK : THE SINGLE VARIABLES MUST BE
C                                       SET BY THE USER; THE DIFFERENT
C                                       RATIOS ARE AUTOMATICALLY
C                                       CALCULATED BY THE COMPUTER.

C     2) VECTOR ID(100) :      THIS VECTOR CONTAINS THE IDENTIFI-
C                              CATION NUMBERS OF ALL SUBJECTS.

C     3) MATRIX IRA(2,50) :    THE FIRST ROW CONTAINS ALL THE
C                              NUMBERS (1,2,.....IV) OF THE
C                              NUMERATORS OF THE DIFFERENT RATIOS.
C                              THE SECOND ROW CONTAINS ALL THE
C                              NUMBERS OF THE DENOMINATORS OF THE
C                              DIFFERENT RATIOS.

C  C) MATRIX SET BY COMPUTER (COMMON BLOCK 'BL')
C  --------------------------------------------
C     MATRIX RMED (4,100) :    THE FIRST ROW CONTAINS THE MEAN
C                              VALUES OF ALL VARIABLES AND RATIOS.
C                              THE SECOND ROW CONTAINS THE STANDARD
C                              DEVIATIONS OF ALL VARIABLES AND RATIOS.
C                              THE THIRD ROW CONTAINS THE VALUES
C                              (XV-TOL*DV) AND (XR-TOLR*DR) OF ALL
C                              VARIABLES AND RATIOS.
```

```
                                THE FOURTH ROW CONTAINS THE VALUES
                                (XV - TOL*DV) AND (XR - TOLR*DR) OF ALL
                                VARIABLES AND RATIOS.

      K1 = IV + 1
      KV = IV + IP
      DO 2 J = K1,KV
      DO 2 I = 1,IR
      IF (RM(I,IRA(2,J-IV)).EQ.0.) GO TO 2
      RM(I,J) = RM(I,IRA(1,J-IV)) / RM(I,IRA(2,J-IV))
    2 CONTINUE
      DO 4 J = 1,KV
      IT = 0
      DO 6 I = 1,IR
      IF (RM(I,J).EQ.0.) GO TO 6
      IT = IT + 1
      RMED(1,J) = RMED(1,J) + RM(I,J)
    6 CONTINUE
      IF (IT.EQ.0) GO TO 8
      RMED(1,J) = RMED(1,J) / IT
      GO TO 10
    8 DO 12 I = 1,4
      RMED(I,J) = 0.
   12 CONTINUE
      GO TO 4
   10 DO 14 I = 1,IR
      IF (RM(I,J).EQ.0) GO TO 14
      RMED(2,J) = RMED(2,J) + (RM(I,J) - RMED(1,J))**2
   14 CONTINUE
      RMED(2,J) = SQRT(RMED(2,J) / (IT-1))
      TOLE = TOL
      IF (J.GT.IV) TOLE = TOLR
      RMED(3,J) = RMED(1,J) - TOLE*RMED(2,J)
      RMED(4,J) = RMED(1,J) + TOLE*RMED(2,J)
    4 CONTINUE
      PRINT 100
  100 FORMAT(16X,21HIDENTIFICATION NUMBER,4X,11HVARIABLE(S),8X,
     *          10HREAL VALUE,11X,10HMEAN VALUE,6X,19HZ-SCORE(ABS. VALUE)
     *          16X,20(1H-),5X,11(1H-),8X,10(1H-),11X,10(1H-),6X,19(1H-))
      DO 16 I = 1,IR
      I2 = 0
      DO 18 J = 1,KV
      IF ((RM(I,J).LE.RMED(4,J).AND.RM(I,J).GE.RMED(3,J)).OR.RM(I,J).
     *    EQ.0) GO TO 18
      I2 = I2 + 1
      Z = ABS((RM(I,J)-RMED(1,J)) / RMED(2,J))
      IF (I2.EQ.1) GO TO 20
      IF (J.LE.IV) GO TO 22
      PRINT 200, IRA(1,J-IV), IRA(2,J-IV),RM(I,J),RMED(1,J),Z
  200 FORMAT(43X,1H(,I2,1H,,I2,1H),8X,F9.3,13X,F9.3,14X,F6.3)
      GO TO 18
   22 PRINT 201,J,RM(I,J),RMED(1,J),Z
  201 FORMAT(43X,1H(,I2,1H),11X,F9.3,13X,F9.3,14X,F6.3)
      GO TO 18
   20 IF (J.LE.IV) GO TO 24
      PRINT 202,ID(I),IRA(1,J-IV),IRA(2,J-IV),RM(I,J),RMED(1,J),Z
  202 FORMAT(22X,I4,17X,1H(,I2,1H,,I2,1H),8X,F9.3,13X,F9.3,14X,F6.3)
      GO TO 18
   24 PRINT 203,ID(I),J,RM(I,J),RMED(1,J),Z
  203 FORMAT(22X,I4,17X,1H(,I2,1H),11X,F9.3,13X,F9.3,14X,F6.3)
   18 CONTINUE
   16 CONTINUE
      RETURN
      END
```

Best Available Document

If, after the second screening, a value was not considered erroneous, the subject numbe was eliminated from the list, and the data sheet was classified unaltered. If, on the othe hand, the value was considered erroneous, the subject number was eliminated from th list, and the correction was noted on the data sheet. Either the value was changed t zero if no alternative or if too many alternatives existed, or it was changed to a nev value if only one clear alternative was possible. As a rule, a 'certain error' was neve replaced by a 'possible real value'.

Third screening: control with available witnesses

In a third screening, the last remaining category of extreme values was examined This was where appropriate witnesses were available which provided several possibl treatments. These were as follows:

(*a*) The witness provided a solution to the problem. If the witness confirmed the value, no correction was needed; the subject's number was removed from the list, and all his information was classified. If, on the other hand, the witness showed the value to be incorrect, the new value was noted on the data sheet.

(*b*) The witness could not give a solution. If the extreme value was *undoubtedly* erroneous, but no alternative value or too many alternative values existed, the extreme value was changed to zero on the data sheet. When only one alternative was possible, the extreme value was adapted on the data sheet. If the extreme value was *perhaps* erroneous, then the same procedure as in (*d*) of the second screening was executed.

Correction of erroneous extreme values

At this point, all the remaining data sheets contained sometimes one or more new values, in many cases one or more zeros and in some cases the label: "to be punched again". The last step of the procedure was to correct the data tape.

4. Example of corrections of some individual cases

The different steps of the correction procedure will now be illustrated by means of four real cases in which several kinds of error were found. Table 1 contains the output obtained from step 4 (Execution of the program for detecting extreme values) for the subjects with identification numbers 0922, 1373, 2147 and 3417.

First screening

In the first screening it was found for subject 0922 that five consecutive variables showed extreme values. The numbers 9 to 13 represent here respectively the variables biacromial diameter, biiliac diameter, biepicondylar femur width, head perimeter and relaxed upper arm perimeter. It is obvious from table 1 that the first four extreme file values for this subject each correspond closely to the calculated mean value of the next variable. The observed file value 21·2 for variable 9 is suspiciously close to the mean value 20·377 of variable 10, while the observed value of 7·6 for variable 10 is much closer to the calculated mean value 8·154 of variable 11. The observed value of 52·1 for variable 11 is close to the calculated mean value of the next variable, and so on. The same five variables also occurred each in at least one extreme ratio. Variables 1, 14, 18 in the ratios represent body weight, thorax perimeter and head length respectively. Obviously this situation resulted either from deleting a variable (probably variable 9) while punching, or from not measuring the variable, or from not recording the value o

ie 1. Exa

Identificati number
0922
1373
2147
3417

ne data sh
suspects'.
Subjec
subscapula
with anthr
method is
between b
nature hig
ios betv
riables 1
extreme va
result of th
data, and
This exam
low eccen
Case 2
53·5 cm c
extreme r
roneous.
he identi
data sheet
found to b
program.
The he
and therel
accompan
the result
sheet. Thi

Best Available Document

101

Table 1. Example of an output obtained from step 4 (Execution of the program for detecting extreme values) for the subjects with identification number 0922, 1373, 2147 and 3417.

Identification number	Variable number	File value	Mean	Percentage deviation
0922	9	21·200	29·222	−27%
	10	7·600	20·377	−62%
	11	52·100	8·154	538%
	12	20·500	52·875	−61%
	13	3·500	19·089	−81%
	1, 13	9·714	1·727	462%
	13, 14	0·058	0·297	−80%
	12, 18	1·133	2·976	−61%
	1, 9	1·604	1·086	47%
	10, 11	0·146	2·671	−94%
	10, 14	0·127	0·317	−60%
1373	16, 17	2·167	1·102	96%
2147	9	53·5	28·9	85%
	4, 5	1·838	1·450	26%
3417	12	5·200	53·095	−90%
	12, 18	0·287	2·906	−90%

the data sheet. As a result of this first screening, subject 0922 was retained on the list of 'suspects', and the original data sheet was requested for the second screening later on.

Subject 1373 showed an extreme ratio of the variables 16 and 17, which are subscapular and suprailiac skinfold. This case is an example of the fact that experience with anthropometric variables is a prerequisite when the presented data cleaning method is used. Relations between skinfolds are of a different nature from, e.g., relations between bone measures. The relative variation of skinfolds and of their ratios is by ture higher than in most other variables. This means that the tolerance for extreme ios between skinfolds should also be higher. Furthermore, the eccentricity of variables 16 or 17, as found in their ratio is neither reinforced by the occurrence of an extreme value of one or both variables nor by another extreme and related ratio. As a result of this first screening, it was decided that this subject's file showed no erroneous data, and the identification number 1373 was deleted from our listings of 'suspects'. This example also illustrates the inadequacy of our system of percentage deviations: the low eccentricity would have been better described in standard deviation units.

Case 2147 showed a highly extreme value for variable 9 (biacromial diameter) of 53·5 cm compared to the mean. The extremeness was not reinforced by a related treme ratio but was nevertheless large enough to be recognized immediately as roneous. Again a measuring or a recording or a punching error could have occurred. The identification number was thus held on the list of extreme values, and the original data sheet was requested. The ratio hand width over wrist width (variables 4 and 5) was found to be not very extreme, and no other extreme related values were indicated by the program. The indication "4, 5" was thus deleted from the list of extreme values.

The head perimeter (variable 12) of subject 3417 was punched on tape as 05·2 cm, and therefore indicated on the listing as extreme. This certain error was of course accompanied by the extreme ratio of head perimeter over head length. If the error was the result of wrong punching, then the correct value could be found on the original sheet. This data sheet was thus requested.

102

Second screening

Examination of the data form of subject 0922 provided no immediate solution, as no punching errors were found. The strong suspicion that variable 9 was never measured, and that four data values had to be moved upwards on the list now had to be checked. Therefore, the subject's somatotype photograph was requested, and subject 0922 remained on the list for the last screening.

The data sheet of subject 2147 showed also the erroneous value of 53·5 cm for biacromial diameter. The value could have been correctly taken but wrongly read from the anthropometer scale, or correctly dictated as 33·5 or 35·5 or even 35·3, but wrongly recorded. No immediate conclusion was possible, so the somatotype photograph was requested as a witness.

The data form of subject 3417 contained the same erroneous value of 05·2 cm for head perimeter as on the data tape. It was not possible to find an appropriate witness to check a perimeter value. Further, the real value could have been 50·2 cm, but alternatively 52·0 cm or even 55·2 cm. Since no witness could be used, since the value was certainly erroneous and since more than one alternative was possible, the wrong value of 05·2 cm was changed to zero on the data sheet. All the corrected forms were held apart for correction of the data tape after the third screening.

Third screening

The measurements on the somatotype photograph of subject 0922 supported completely the hypothesis of the missing variable 9. Indeed, the measurement on the photograph indicated that the real values of variables 10 and 11 were close to 21·2 and 7·6 respectively. On the other hand, the real value of variable 9 was close to 30·0. A value of 21·2 was impossible. As a result, the value for biacromial diameter was brought to zero and the next four variables received their exact data value on the subject's data sheet. This form was held apart for later correction of the data tape.

The biacromial diameter and other body measurements were measured on the somatotype photograph of subject 2147, and compared to the possible alternatives for 53·5 cm. We were able to decide on the exactness of the 33·5 value. The correction was noted on the data sheet, which was held apart for correction of the data tape after the third screening of all the remaining 'suspect' variables.

5. Results

The following observations can be made about the process. When the detection and correction of data errors was carried out, it was seen that almost no punching errors occurred. This is probably due to the very safe 'double punching' that had been used originally. However, the listing of subjects classified according to the data on tape could not be trusted. Instead, the complete listing of the data tape was used. Errors of classification and numbering were detected on the tape that contained all the subjects of the study. Some of these errors would not have been detected if the program had been run from a tape organized into different subject groups, arranged according to age, language, sex, etc.

The following kinds of basic error were detected on the original data sheet:

1. Normal measuring errors (e.g. 29·6 instead of 19·6)
2. Wrong order of measuring, particularly when measuring instruments were changed like changing from a big to a small spreading caliper.
3. Transposition of numerals (82·3 instead of 28·3; 28·0 instead of 20·8).

4. Results dictated incompletely (9·5 instead of 09·5).
5. Phonetic errors, for example: 9 instead of 5, or 80 instead of 18).
6. Key punch errors which almost always resulted from bad handwriting.

In some cases, an error was detected through only one extreme ratio value, if this value was really impossible (e.g. a ratio sitting height/tibia length smaller than 1).

As some errors cause shifts of the values on the data sheet, it was sometimes possible to correct several values on one data sheet. If the direction of the shift was obvious, this often led to the correction of not extreme, but erroneous values, where the error had been obscured by the shift.

Certain aids proved to be invaluable in the process of detection. Among these, frequency distributions of the single variables were indispensable for the decisions about the eccentricity of values. In addition, somatotype photographs were often used to check maturity codes.

Finally, if supporting data were available, they usually provided reliable information. If they were needed but not available, the correction procedure was rather tedious and the chances of correcting data decreased considerably.

Table 2 quantifies the elaboration of the working procedure. In the preparatory step, 49 errors in the classification of the data were corrected. From the 1986 subjects with extreme values detected with the specific program, 1585 were considered to be

Table 2. Summary of corrections.

Number of subjects in the study: 6969			Corrected files	Corrected errors
Correction of classification errors				
Number of files corrected: 49			49	
First screening				
Files containing one or more extreme values detected:		1986		
Those files containing extreme but correct values:		−1585		
Those files containing possible errors:		401		
Second screening				
Data files controlled:		401		
Files containing extreme but correct values		−118		
Files corrected:		−99	99	
(errors corrected: to zero :	51			51
to new value :	123)			123
Witnesses available:		184		
Third screening				
Witnesses controlled : photographs :	164			
X-rays :	13	184		
footprints :	7			
Files containing extreme but correct values :		−23		
Files corrected :		161	161	
(errors corrected : to zero :	123			123
to new value :	73)			73
Total number of files corrected :			309	
Total number of errors corrected :				370

104

extreme but correct. The remaining 401 'suspects' were then checked against their original data sheets, and 118 of them were considered to be correct. Many of the 283 remaining files containing possible errors could not be corrected immediately. If in these no witnesses were available, the erroneous value was changed to zero. In 184 cases witnesses were available and were thus requested. At the same time, the certain errors, where no witness was needed, were also changed to their new value or to zero.

In the third screening, the remaining 184 cases were checked against their witnesses. Twenty-three were extreme, but correct. The other 161 were corrected in the most appropriate way. This brought the total number of subjects for which corrections were made to 309 but the total number of corrections was 370.

6. Discussion

Table 2 indicates not only the total number of corrected errors, but also the total number of corrected subject files. Indeed, in many cases, one error caused more in the same subject's file. Seen in the light of corrected against existing files, the number of corrections was approximately 4·4%. If the real number of corrected errors (which is 370) is compared to the total number of subjects, this percentage reaches 5·5%. This means that no less than one error for every 18 subjects could be identified and corrected or altered to zero. The zero is regarded as a missing value.

The large number of detected files with extreme values that were obtained after the first screening (1986) gives reason to believe that few recognizable or detectable errors were ignored. Furthermore, the number of cases rejected from the possible errors in this first screening (1585) guarantees that the true extreme data values were respected.

The limits that were used, to decide whether or not a suspected variable or ratio was extreme, are subject to discussion. These limits depend on the specific range of each variable and each ratio separately. The ultimate decision was taken only after careful consideration of the complete combination of all of the subject's measurements.

The percentage deviations of extreme values from the group mean for the same variable were often misleading, because they did not take into account the specific variance of the variable. Absolute values of the z-scores are now used in our error correction procedures and the program contains this modification. A subroutine to detect systematically the existence of really impossible ratios (e.g. the ratio standing height/sitting height smaller than 1) will also be included in further error detection programs. If data from longitudinal studies have to be corrected, then additional correction procedures will need to be devised. As many biological characteristics are not reversible, the control of this feature will constitute an additional criterion for error detection. Comparing each value with the time series of the individual for the same variable will also be an important feature.

The recording of growth study data potentially involves many errors. It is clear that most of these errors could be avoided if the data recording were automated. Specially designed instruments exist for this purpose and have already been used in anthropometric studies (Prahl-Andersen *et al.* 1972). In circumstances where automated recording is impossible, the collected data should be recorded on specially prepared sheets by a trained recorder. This recorder should be a member of the measuring team, who is acquainted with the measures, their means and ranges. This means that information should not be recorded by one of the subjects in the study. Furthermore, if the measured values are dictated to another person, a pre-determined system should be followed and strictly respected. A last possible resource, if no automated instruments or no reliable recorders are available, is to dictate all measured values to a tape recorder.

105

It is also clear that the number of data errors caused by, for instance, wrong recording will be substantially reduced, if the measurements are duplicated on the spot either by the same or by a different investigator. However, this is rather a question of validity of data.

Acknowledgements

This study is part of the research project *Performance and Talent*. This investigation was supported by the Belgian Ministry of Education and Culture (BLOSO & ADEPS), the Ministry of Health and Family (CBGS), and the Belgian Fund for Fundamental Collective Research (FKFO) under contract no. 935. It is also part of the research project *LLEGS* (Leuven Longitudinal Ecologic and Experimental Growth Study), which is supported by the Belgian Fund for Medical Research (FGWO) under contract no. 3.0047.75.

References

HEBBELINCK, M., and CLIQUET, R. L., 1970, "Performance and Talent". An anthropological and sociological investigation of primary school children. *Gymnasion*, 7, 7–15.

PRAHL-ANDERSEN, B., POLLMANN, A. J., RAABEN, D. J., and PETERS, K. A., 1972, Automated anthropometry. *American Journal of Physical Anthropology*, 37, 151–154.

Address correspondence to: W. Duquet, Vrije Universiteit Brussel, Hilok Laboratory of Human Biometry and Biomechanics, Pleinlaan 2, B-1050 Brussel, Belgium.

Zusammenfassung. Diese Arbeit beschreibt eine neue Methode zur Datenglättung bei Wachstums- und Entwicklung-Studien. Die Methode basiert im wesentlichen auf der Annahme, daß ein Extremwert einer bestimmten Variablen verdächtig sein könnte, wenn andere, hochkorrelierte Variable wenig zu ihm passen, und daß auf der anderen Seite ein extremer, aber möglicher Wert wohl eher nicht als falsch anzusehen ist, wenn er durch einen anderen extremen, aber möglichen Wert einer hochkorrelierten Variablen bestärkt wird. Jede Variable wurde daher in mindestens zwei Verhältniszahlen mit den höchstkorrelierenden Variablen verwendet. Es wurde ein Programm entwickelt, um Extremwerte von Einzelvariablen und extreme Verhältniswerte aufzuspüren. Das Vorgehen mit Verhältnissen wurde als Hilfe benutzt, um mögliche Fehler entdecken und sie von wahren Extremwerten zu unterscheiden. Beim Vergleich zwischen korrigierten und ursprünglichen Daten betrug die Häufigkeit von Korrekturen etwa 4,4%. Wenn die tatsächliche Zahl korrigierter Fehler mit der Gesamtzahl der Probanden verglichen wird, steigt dieser Prozentsatz auf 5,5 an. Wenn bei der Korrektur der wirkliche Wert nicht mit Sicherheit gefunden werden konnte, wurde der falsche Wert einfach ausgelassen. Aus unseren Ergebnissen zeigt sich, daß man Grund hat zur Annahme, daß dieser Methode wenige entdeckbare Irrtümer dem Glättungsvorgang entgehen. Weiterhin werden in dieser Arbeit Schlußfolgerungen für zukünftige Korrekturmaßnahmen und für zukünftige Wachstumsstudien im allgemeinen gegeben.

Résumé. Le travail décrit une nouvelle méthode pour lisser les données dans les études de croissance et développement. La méthode est basée essentiellement sur la supposition qu'une valeur extrême d'une certaine variable pourrait être douteuse si d'autres variables hautement corrélées avec elle montrent une faible compatibilité avec elle et, d'un autre côté, qu'une valeur extrême mais possible tend à ne pas être fautive si elle est renforcée par une valeur extrême mais possible d'une autre variable hautement corrélée. Chaque variable a été dès lors inclue dans au moins deux quotients avec les variables les plus corrélées. Un programme a été établi pour déceler les valeurs extrêmes des variables individuelles et des quotients. Cette procédure avec quotients a été employée pour déceler les erreurs possibles et les discriminer de valeurs extrêmes vraies. A la lumière du fichier de données corrigées par rapport au fichier de données originales, le nombre de corrections était d'environ 4,4%. Si le nombre réel d'erreurs corrigées est comparé au nombre total de sujets, ce pourcentage atteint 5,5%. Si, lors de la correction, la valeur réelle n'était pas décelée avec certitude, la valeur erronée était simplement écartée. De nos résultats, il y a des raisons de croire qu'avec cette méthode peu d'erreurs décelables échapperont à la procédure de lissage. Des conclusions sur des procédures futures de correction et sur des études futures de croissance en général sont aussi données dans ce travail.

106

STATISTICS IN MEDICINE, VOL. 4, 457–467 (1985)

THE USE OF AN AUTOREGRESSIVE MODEL FOR THE ANALYSIS OF LONGITUDINAL DATA IN EPIDEMIOLOGIC STUDIES

BERNARD ROSNER

Channing Laboratory and the Department of Preventive Medicine and Clinical Epidemiology, Harvard Medical School, 180 Longwood Avenue, Boston, MA 02115, U.S.A.

ALVARO MUÑOZ

Channing Laboratory and the Department of Medicine, Harvard Medical School, 180 Longwood Avenue, Boston, MA 02115, U.S.A.

IRA TAGER, FRANK SPEIZER AND SCOTT WEISS

Channing Laboratory and the Joint Department of Medicine, Beth Israel and Brigham and Women's Hospital, Harvard Medical School, 180 Longwood Avenue, Boston, MA 02115, U.S.A.

SUMMARY

Korn and Whittemore[1,2] have presented methods for analyzing longitudinal data where the number of observations per individual is large relative to the number of variables considered for each subject. However, this is often not the case in epidemiologic studies, since one usually collects data at relatively few time points, and the quantity of data collected for each individual at each time point is typically extensive. We present here an autoregressive model for analyzing longitudinal data of this type for the case of a continuous outcome variable. Some of the important features of this model are that one can (1) in the same analysis, consider both independent variables that are time-dependent and those that are fixed over time, (2) partially use data for an individual where some examinations are missing, (3) assess relationships between changes in outcome and exposure over short periods of time, (4) use ordinary multiple regression methods. Anderson[3] has considered this type of model, but, to our knowledge, the model has never been applied to biostatistical problems. We illustrate these methods with data from a longitudinal study that seeks to identify the role of personal cigarette smoking on changes in pulmonary function in children.

KEY WORDS Longitudinal data Autoregressive time series Pulmonary function data Regression methods

INTRODUCTION

Longitudinal epidemiologic studies often involve collection of both exposure and outcome information at several points in time. In such studies, for any one individual, the number of exposure variables studied at each time point is often large, while the number of time points for collection of information is small (such as in the case of annual followup examinations). In addition, a considerable number of individuals may lack observation at several points in time and the resulting missing data problem becomes an important issue.

Interest often focuses on the relation between changes in exposure to changes in outcome over time. The analytic approach involving cross-sectional data analyses at each point in time is inherently inadequate for dealing with the above problem. Another commonly used method is to

0277–6715/85/040457–11$01.10
© 1985 by John Wiley & Sons, Ltd.

Received November 1984
Revised February 1985

107

relate differences in exposure directly to differences in outcome over two time points. Rosner[4] has discussed the inadequacies of this method. Korn and Whittemore[1,2] have considered a similar problem for a binary outcome where, for a given person, the number of time points with information available is large relative to the number of exposure variables considered. The same authors[5] have also considered this problem for continuous outcome variables.

In this paper, we present a method for analyzing longitudinal data for the case of a continuous outcome variable with examinations equally spaced over time where the number of time points is small relative to the number of exposure variables. This method (1) uses independent variables for an individual that both do and do not change over time (such as height and sex, respectively, for children), (2) does not require complete data on all individuals, but can partially use information from individuals with some examinations missing, and (3) allows for use of ordinary least squares methods available in standard statistical packages such as SAS or BMDP upon construction of the appropriate derived data file.

The models presented here have had frequent use in econometric modelling[3], but to our knowledge, have never been applied in epidemiologic research. We illustrate these methods with data collected annually for four years in a study of the effect of personal cigarette smoking on changes in pulmonary function in children.

METHODS

General model

Given T longitudinal examinations, we model the outcome of the ith subject at time t as a normal variate whose mean is a linear function of the outcomes at previous time points as well as the exposure variables ascertained at time t, and whose conditional variance is the same over time and across individuals. Specifically,

$$y_{it} = \alpha + \sum_{l=1}^{L} \gamma_l y_{i,t-l} + \sum_{j=1}^{J} \beta_j x_{ijt} + \sum_{k=1}^{K} \beta_k^* z_{ik} + e_{it} \tag{1}$$

where $i = 1, \ldots, n$; $t = L, \ldots, T$; y_{it} = value of the outcome variable for the ith individual at the tth examination, e_{it} are statistically independent for all i, t with a common $N(0, \sigma^2)$ distribution. In the context of econometric applications, Anderson[3] refers to the model in equation (1) as an Lth order autoregressive model with 'independent variables'.

In equation (1), the x's represent exposure variables that change over time, such as height and smoking status; thus, x_{ijt} is the jth time-dependent exposure variable for the ith subject ascertained at time t. We note that the x's, although indexed at time t, could also represent changes in exposure between several time points, such as the change in height between successive exams, or more complicated functions of exposure over several time points. The z's represent exposure variables that do not change over time such as sex and race; thus, z_{ik} is the kth fixed exposure variable for the ith subject. The γ's represent the effect of the previous y's on the current level of y, while the β's and β^*'s represent the effect of the independent variables on the level of the outcome variable at time t after adjusting for levels of the outcome variable at the previous L time points. In particular, if $L = 1$, then the β's and β^*'s represent the effect of the independent variables on level at time t after adjusting for level at time $t - 1$.

The model in equation (1) makes the following assumptions:

(a) the residuals e_{it} are independent with constant variance σ^2, both for data obtained at different time points within the same individual, and for different individuals.

158

(b) the same relationship between outcome and exposure exists for different individuals under study, i.e. we assume a fixed effects model. It is this assumption which allows precise estimation of the coefficients of the model.

The model in equation (1) has the feature that we can, in a unified fashion, consider variables that change over time $\{x_{ijt}\}$ as well as those that do not $\{z_{ik}\}$. This is in contrast to Korn and Whittemore's[5] approach that involves, at a first stage, regression modelling performed separately for each individual person based on the time-dependent covariates only, followed by, at a second stage, averaging the individual regression coefficients according to the values of the fixed covariates (i.e. sex, race, etc.). Laird and Ware[6] have extended this approach by performing both stages in the context of a random effects model and with use of empirical Bayes techniques. One disadvantage of the two-stage models is that they often require specialized software, while standard regression packages can be used to fit the model in equation (1). Another disadvantage of two-stage models is that the number of time-dependent covariates considered for an individual must be smaller than the number of visits.

Individuals with no missing data contribute $T-L$ rows to the design matrix represented by the regression model in equation (1). Longitudinal studies, however, often entail missing data for an individual on one or more exams. The model in equation (1) is stated explicitly for complete data and a modification allowing for missing data would be useful in practice. We propose to use the complete case method for the treatment of missing data[7], namely, an individual contributes an observation to the model at time t, if all variables used in the model are present. Due to the time-dependent nature of some of the independent variables, this usually will require the presence of data at some but not all of the previous time points. This implies that each individual will contribute a different number of observations to the model depending on how much data are present. In particular, we can include information for any individual for all collections of time points $\{t, t-1, \ldots, t-L\}$ where information is available for all relevant variables. Thus, for example, if $L = 1, T = 7$, and an individual is not present for exam 3, then that individual can still contribute information to the model in equation (1) for the pairs $(t, t-1) = (2, 1), (5, 4), (6, 5), (7, 6)$. This information can be handled with the use of ordinary least squares regression methods without any special missing data algorithms upon construction of a derived data file indexed by available pairs of visits for a subject rather than by the subjects themselves.

An issue in fitting models of this type is the determination of the appropriate value of L. To determine L, the model fitting problem can be considered as a multiple regression problem in which one is testing for the significance of a specific set of covariates (A_1) after controlling for previously known covariates (A_0) and previous values of the outcome variable. The partitioning of covariates into the subsets A_0 and A_1 is usually based on pre-existing data: the important issue as regards determination of L is use of the same subset A_0 for all values of L considered. Given the subset A_0, we use stepwise regression methods as given in Anderson[3] to select the appropriate value for L. In the same way, given A_0 and L, we determine the significance of the covariates of interest (A_1) by similar methods. Thus, the model fitting begins by initially identifying a set of nuisance parameters that must always be present in the model (A_0), and proceeds in a two-stage stepwise regression process to determine (a) the appropriate value of L given A_0 (b) the significance of potential risk factors in A_1 given A_0 and L. We note that the hypothesis tests used in (a) to determine the appropriate value of L are based on successively smaller data sets since the number of rows in the design matrix decreases as L increases. Furthermore, with L determined as $T-1$, the order of the process may not be well estimated, since the optimal L might be larger than $T-1$ if more examinations were available for an individual.

The model in equation (1) assumes (a) uncorrelated residuals for the various exams within an

109

individual (b) constancy of effects across individuals, i.e. fixed effects. A method for checking assumption (a) is to generalize model 1 by allowing the residuals for an individual to follow an intraclass correlation structure[8] and to test whether the intraclass correlation is significantly greater than 0. A method for fitting this generalized model in the presence of covariates is given in Rosner[9] and a SAS procedure is available from the authors to implement this method. As regards assumption (b), for reasons stated previously, it is difficult to fit a full random effects model. A compromise solution is to introduce terms of the form $z_{ik}x_{ijt}$ to equation (1) which will allow one to test whether the effects of the jth time-dependent covariate are the same in different subgroups defined by the person-specific covariate z_{ik}.

Statistical characteristics of the model

We focus here on certain statistical properties of the model in equation (1) in the case of $L = J = K = 1$, i.e.

$$y_{it} = \alpha + \gamma y_{i,t-1} + \beta x_{it} + \beta^* z_i + e_{it} \tag{2}$$

where, for ease of notation, we have replaced $\gamma_1, \beta_1, \beta_1^*$ by γ, β, β^* respectively. This represents the conditional distribution of y_{it} given $y_{i,t-1}, x_{it}$, and z_i. Using equation (2) recursively, we can remove the effects of all previous y's. This process yields the conditional distribution of y_{it} given the x's and z's as:

$$\begin{bmatrix} y_{i0} \\ y_{i1} \\ \vdots \\ y_{it} \\ \vdots \\ y_{iT} \end{bmatrix} = \begin{bmatrix} \alpha_0 \\ \gamma\alpha_0 + \alpha \\ \vdots \\ \gamma^t\alpha_0 + \dfrac{\alpha(1-\gamma^t)}{(1-\gamma)} \\ \vdots \end{bmatrix} + \begin{bmatrix} \beta_0 & 0 & \dots & & & & & 0 \\ \gamma\beta_0 & \beta & 0 & \dots & & & & 0 \\ \vdots & & & & & & & \\ \gamma^t\beta_0 & \gamma^{t-1}\beta & \dots & \gamma\beta & \beta & 0 & \dots & 0 \\ \vdots & & & \vdots & & & & \vdots \end{bmatrix} \begin{bmatrix} x_{i0} \\ x_{i1} \\ \vdots \\ x_{it} \\ \vdots \\ x_{iT} \end{bmatrix}$$

$$+ \begin{bmatrix} \beta_0^* \\ \gamma\beta_0^* + \beta^* \\ \vdots \\ \gamma^t\beta_0^* + \dfrac{\beta^*(1-\gamma^t)}{(1-\gamma)} \end{bmatrix} z_i + \begin{bmatrix} \varepsilon_{i0} \\ \varepsilon_{i1} \\ \vdots \\ \varepsilon_{it} \\ \vdots \\ \varepsilon_{iT} \end{bmatrix} \tag{3}$$

where $\alpha_0, \beta_0, \beta_0^*$ are the parameters of the conditional distribution of y_{i0} given x_{i0}, z_i (i.e. the cross-sectional distribution of y_{i0}) and $\varepsilon \sim N(0, \Sigma)$ where $\mathrm{var}(\varepsilon_{it}) = \sigma^2/(1-\gamma^2)$ and $\mathrm{corr}(\varepsilon_{it}, \varepsilon_{it'}) = \gamma^{|t-t'|}$. We see from equation (3) that the expected value of y_{it} depends on *all* previous values of x with regression coefficient $\gamma^{t-s}\beta$ for x_{is} $(0 \leqslant s \leqslant t)$. Thus, x's at times close to t receive more weight than x's further away from t. This is in contrast to standard autoregressive modelling where

$$y_{it} = \alpha_1 + \beta_1 x_{it} + \beta_1^* z_i + u_{it}$$

and

$$u_{it} = \gamma u_{i,t-1} + a_{it}, \tag{4}$$

where a_{it} are i.i.d. $N(0, \sigma_1^2)$.

The principal difference between models (3) and (4) is the expression for the expected value of y and hence in the resulting interpretation of the regression coefficients of x_{it} and z_i. In particular, in (4) y_{it} only depends on x_{it} and z_i while in (3) y_{it} depends on z_i and on all previous values of x with

110

successively less weight given to x's further away from t. Model (3) is more reasonable since previous values of x influence previous values of y which in turn relate to current values of y. Furthermore, the model in (4) can be rewritten in the following form

$$y_{it} = \alpha_1(1-\gamma) + \gamma y_{i,t-1} + \beta_1 x_{it} - \beta_1 \gamma x_{i,t-1} + \beta_1^*(1-\gamma)z_i + a_{it}. \tag{5}$$

This is a special case of the general model in equation (1) with the coefficient of $x_{i,t-1}$ constrained to be the negative of the product of the coefficients of x_{it} and $y_{i,t-1}$.

In the case that the distribution of y_{it} does not depend on the previous y's, the model in equation (2) reduces to a cross-sectional analysis based on treating data provided by an individual at multiple time points as independent observations.

We note that we can also use the model in equation (2) to estimate the effects of individual covariates over more than one time period. In particular, assume that for two hypothetical individuals (i, i'), $z_i = z_i'$ and $x_{it} = x_{i't}$ for $t \leqslant t_0$ and $x_{it} = x_{i't} + \Delta$ for $t > t_0$. We can then show, using equation (3), that

$$E(y_{i,t_0+\tau} - y_{i',t_0+\tau}) = \beta\Delta(1-\gamma^\tau)/(1-\gamma) \tag{6}$$

where $\tau \geqslant 1$ and E represents the expected value. Thus, with use of equation (6), we can quantify the effects of particular covariates on outcome over time after holding all other covariates fixed. Furthermore, the results of equation (6) hold for more than one fixed or time-dependent covariate, provided that (a) the levels of all fixed covariates are the same for both individuals and (b) the levels of all time-dependent covariates other than the variable of interest are the same over all time points.

Results similar to those given in equation (6) hold by allowing the fixed covariate for two hypothetical individuals to differ by Δ^*, with the assumption that the time dependent covariates are the same for both individuals over all time points and that the other fixed covariates if present are the same for the two individuals. In this case, we replace β by β^* and Δ by Δ^* in equation (6).

We see from equation (3) that the partial correlation between an outcome variable assessed at two different points in time is assumed to be the same for all individuals. In many instances, this may not be the case. For example, in pulmonary function testing, subjects with respiratory symptoms may have more unstable levels of pulmonary function over time than asymptomatic individuals. We can easily accomodate this situation by specifying different values of γ for the two different groups of individuals. We can accomplish this by including an interaction term of the form $y_{i,t-1}z$ where z is an indicator variable for membership in a particular group. We can generalize this if g groups of individuals have different levels of γ by introducing $g-1$ such indicator variables.

EXAMPLE

We consider an example of the above model with use of pulmonary function data collected annually over seven years in children ages 6–19 in East Boston, Massachusetts. These children are part of a longitudinal study of early life risk predictors for obstructive airways disease[10]. Information concerning personal smoking of these children was obtained from the children themselves, in the absence of their parents, using standardized questionnaires[11]. Measures of forced expiratory volume were obtained with an 8-liter, water-filled spirometer in the sitting position without noseclip. Forced expiratory volume 1-second (FEV_1) was obtained by standard methods[12].

For the purposes of this example, we focus on data obtained in years 4–7 of the study, since personal smoking information was available for all subjects in the above age range only in years 1, 4, 5, 6, and 7. We did not use data from year 1 since the methods of this paper require pairs of observations with complete data in successive equally spaced exams (see equation (1)). The detailed

Table I. Pattern of visits with complete data over four years, children 6–19 years of age, East Boston, MA, 1977–1980

	Year				Number of children with complete data	Number of pairs of visits used in the analysis per child	Total number of pairs of visits
	4	5	6	7			
	X	X	X	X	395	3	1185
	X	X	X	.	51	2	102
	.	X	X	X	28	2	56
	X	X	.	X	75	1	75
	X	X	.	.	49	1	49
	X	.	X	X	27	1	27
	.	X	X	.	10	1	10
	.	.	X	X	39	1	39
	X	.	X	.	15	0	0
	X	.	.	X	5	0	0
	X	.	.	.	65	0	0
	.	X	.	X	10	0	0
	.	X	.	.	13	0	0
	.	.	X	.	13	0	0
	.	.	.	X	75	0	0
Total	682*	631	578	654	870		1543

X = Present.
. = Missing.
* = Total number of persons at each respective visit.

methods of procedure for this study appear elsewhere[10]. A complete discussion of the epidemiologic findings using the entire data set is presented in Tager *et al.*[13].

A total of 870 children had complete data for FEV_1, height, age, sex, and personal smoking habit in at least one of the four years. Table I gives the pattern of visits with complete data for these individuals. We used data in the analysis from 674 (77·5 per cent) of the 870 individuals who had at least two consecutive visits with complete data. These individuals provided a total of 1543 pairs of consecutive visits. Of these 674 individuals, 395 had complete data on all visits and provided information to the analysis on 1185 (76·8 per cent) pairs of visits, while the remaining 279 individuals had some missing data yet provided 358 (23·2 per cent) pairs of visits to the analysis. If we had used conventional methods of analysis requiring complete data for all visits, we would not have used the latter information. To look at the possible bias introduced by including people in the analysis with some missing data, we have compared the FEV_1 at year 4 between the subgroups of individuals who did and did not provide data in year 5. In particular, a multiple regression analysis was performed with FEV_1 at year 4 as the dependent variable and sex, age, height, current smoking habit at year 4 and an indicator variable for the presence of individual in year 5 as independent variables. The partial regression coefficient of the indicator variable was $-0{\cdot}031 \pm 0{\cdot}039$ ($p = 0{\cdot}43$) which indicates that presence in year 5 was not significantly related to level of FEV_1 at year 4. Similar results were obtained when introducing indicator variables for the number of examinations provided in years 5, 6 and 7. Thus, we conclude that no bias is introduced by combining information from individuals with some missing data with information from individuals with complete data.

112

Table II. Characteristics of study population at each examination, children 6–19 years of age, East Boston, MA 1977–1980

	Year			
	4	5	6	7
Characteristic	n (%)	n (%)	n (%)	n (%)
Age at year four (years)				
6–9	226 (33·1)	219 (34·7)	237 (41·0)	309 (47·2)
10–14	364 (53·4)	336 (53·2)	292 (50·5)	296 (45·3)
15–19	92 (13·5)	76 (12·0)	49 (8·5)	49 (7·5)
Sex				
Male	359 (52·6)	327 (51·8)	301 (52·1)	336 (51·4)
Female	323 (47·4)	304 (48·2)	277 (47·9)	318 (48·6)
Current smoker				
Yes	39 (5·7)	46 (7·3)	42 (7·3)	65 (9·9)
No	643 (94·3)	585 (92·7)	536 (92·7)	589 (90·1)

		Year										
		4			5			6			7		
Characteristic		mean	sd	n	mean	sd	n	mean	sd	n	mean	sd	n
	Age at year four (years)												
Height (cm)	6–9	131·2	7·9	226	137·7	8·2	219	142·2	8·9	237	144·6	11·2	309
	10–14	149·0	10·4	364	155·2	10·2	336	159·9	9·1	292	164·0	9·3	296
	15–19	167·2	8·9	92	169·2	8·6	76	168·1	8·5	49	170·4	8·6	49
FEV (liters)	6–9	1·59	0·28	226	1·75	0·31	219	1·90	0·39	237	2·03	0·51	309
	10–14	2·28	0·56	364	2·56	0·63	336	2·78	0·65	292	3·09	0·70	296
	15–19	3·56	0·75	92	3·64	0·81	76	3·55	0·72	49	3·71	0·88	49

Table II describes the characteristics of the study population at each examination. Approximately 85–90 per cent of the study population is under 15 and about 52 per cent are male at any particular examination. The per cent of current smokers increases from 5·7 per cent at year four to 9·9 per cent at year seven. The rates of growth of both height and FEV_1 vary considerably by age. In particular, the younger individuals exhibited the most pronounced growth in height (age 10–14, 15·0 cm; age 6–9, 13·4 cm; age 15–19, 3·2 cm). Similar trends were obtained for growth in FEV_1 over 3 years (age 10–14, 0·81 liters; age 6–9, 0·44 liters; age 15–19, 0·15 liters). Furthermore, the increase in mean FEV_1 over time in particular age groups also was accompanied by a corresponding increase in the standard deviation.

We now proceed to fit a model of the form given in equation 1 based on the 1543 pairs of visits contributed by 674 individuals. Since it is natural to consider changes in FEV_1 in terms of ratios rather than differences, we use the natural logarithm of FEV_1 in our subsequent model-fitting. In addition, the logarithm better satisfies the linearity and normality assumptions inherent in equation (1). Before considering other covariates we need to determine the appropriate number of

113

Table III. The model with no covariates*

L	n†	r‡	Model parameters	Parameter estimate	Standard error	p-value
1	674	1543	α	0·855		<0·001
			γ_1	0·922	0·0082	<0·001
2	474	869	α	0·857		
			γ_1	0·859	0·0376	
			γ_2	0·058	0·0369	0·119

The two models considered are of the form

$$\ln \mathrm{FEV}_{it} = \alpha + \sum_{l=1}^{L} \gamma_l (\ln \mathrm{FEV}_{i,t-l} - c_l) + e_{it},$$

$$i = 1, \ldots, n, \quad t = 2, 3, 4$$

for $L = 1, 2$ respectively, where c_l are centering constants, $c_1 = 0{\cdot}75$, $c_2 = 0{\cdot}65$, chosen to be close to the median ln (FEV) in years 5 and 4 respectively, thereby simplifying the interpretation of the constant term (α).

† n = number of individuals.

‡ r = number of rows in the design matrix in fitting specific models.

time points (L) to be used in the model. For the purposes of this example, we do not prespecify any covariates in the model before determining L (i.e. A_0 is empty). In Table III, we present the fitted parameters for the model with no covariates given $L = 1$ and 2, respectively. We note that for the case of $L = 2$ we could use only 869 observations (triplets) contributed by 474 individuals in the analysis, such that for each triplet, complete data were available for FEV_1, height, age, sex, and current smoking habit for each of the three years. It follows from Table III that we need consider only the immediately previous time point in the model (i.e. $L = 1$).

In Table IV, we present results concerning the relationship between changes in FEV_1 over time and other covariates of interest, including the time-dependent covariates of age, height, growth, and current smoking habit and the fixed covariate of sex, using the model in equation (1) with $L = 1$.

We have considered two models in our data analyses:

(1) an initial model including effects for previous level of FEV_1, age, sex, height, and growth but excluding the effect of current smoking.

(2) a second model including current smoking in addition to all variables in (a).

Table IV shows that for both models, the effects of previous FEV_1, height, and growth are highly significant predictors of current FEV_1 ($p < 0{\cdot}001$), while sex is an additional significant predictor (model 1, $p = 0{\cdot}019$; model 2, $p = 0{\cdot}025$). Furthermore, we see from comparing models 1 and 2, that current smoking is a significant ($p < 0{\cdot}001$) predictor of FEV_1 after controlling for the former variables. Finally, age, which was a significant ($p = 0{\cdot}049$) predictor in model 1, is no longer significant (model 2, $p = 0{\cdot}710$) after controlling for cigarette smoking.

The direction of the predicted effects from model 2 are reasonable in that previous FEV_1, previous height and rate of growth are positively related to subsequent FEV_1, while current cigarette smoking is negatively related to FEV_1. In addition, males are predicted to have slightly higher growth rates in FEV_1 than females after controlling for age, initial FEV_1, initial height, growth in height, and cigarette smoking.

To check for the assumption of the independence of residuals, we have fitted the model in

Table IV. Relationships between changes in FEV_1 over time and other covariates based on 1543 pairs of observations obtained from 674 individuals

Variable	Model 1			Model 2		
	Regression coefficient	se*	p-value (2-tail)	Regression coefficient	se	p-value (2-tail)
Constant	0·7374			0·7488		
$\ln(FEV_{t-1}) - 0{\cdot}75$	0·7767	0·0158	<0·001	0·7712	0·0158	<0·001
$\ln(\text{height}_{t-1}) - 0{\cdot}33$	0·8050	0·0547	<0·001	0·7872	0·0546	<0·001
$\ln(\text{height}_t/\text{height}_{t-1})$	1·6098	0·0956	<0·001	1·5781	0·0953	<0·001
sex†	0·0105	0·0045	0·019	0·0100	0·0044	0·025
$\text{age}_t - 15{\cdot}0$	−0·0032	0·0016	0·049	−0·0006	0·0017	0·710
$(\text{current smoking})^{\ddagger}_{t-1}$		—		−0·0503	0·0116	<0·001

* Standard error.
† Sex is defined as 1 if male and 0 if female.
‡ Current smoking is defined as 1 if yes and 0 if no.

equation (1) allowing for an intraclass correlation structure between the residuals for an individual. The estimated intraclass correlation between residuals was −0·09 in this model which provides no evidence of a positive intraclass correlation.

We can also use the coefficients in model 2 of Table IV to quantify the effect of personal smoking on a child's respiratory function for periods of longer than one year. For illustration, consider two children with identical age, sex, height, and FEV_1 at some time t_0 who have never smoked up to this time. We assume that at t_0 one of the children becomes a smoker and, henceforth, does not quit, while the other child remains a non-smoker. We also assume that both children show the same growth in height after t_0. Using equation (6) and model 2 in Table IV, the estimate of the ratio of the FEV_1 of the smoking child to the FEV_1 of the non-smoking child after 3 years is $\exp(-0{\cdot}119) = 88{\cdot}8$ per cent. This difference could be substantial by age 20 since, from Table II, a typical FEV_1 for a 20 year-old in this data set is 3·7 litres, and the magnitude of the above effect for an adolescent who commences to smoke at age 17 would be on the order of 0·42 liters $[(1 - 0{\cdot}888) \times 3{\cdot}7]$ by age 20.

DISCUSSION

We have proposed a model for the analysis of longitudinal data and have presented an application of this model in the context of an assessment of factors that affect the growth of pulmonary function in children. This model allows direct assessment of the effects of specific covariates over one time period. Furthermore, using equation (6), we can make predictions of the effects of specific covariates over more than one time period, under the assumption that the process is Markovian ($L = 1$). In addition, for non-Markovian models ($L > 1$), we can, using similar recursive methods, generalize the formula given in equation (6) as a function of β and γ. We note, however, that one must exercise caution in making predictions over long periods of time, particularly time periods longer than the maximum follow-up time in the observed data set.

The model presented in equation (1) has the key assumption that the residuals for a particular individual will be independent after conditioning on outcome at the previous L time points. In some applications, this assumption may be violated, regardless of the value chosen for L. In this case, each individual contributes a single observation to the design matrix $L = T - 1$ and one proceeds as in equation (1). Another important assumption is that observations are equally spaced

over time. This is necessary since the correlation between pairs of unequally spaced exams will not be the same as is assumed in equation (1). This is an important limitation of the application of these methods to general longitudinal data sets, since often, and by design, exams are not equally spaced, e.g. in a clinical trial, there are usually more exams per unit time at the beginning of a trial than at the end. We are now investigating an extension of the model in equation (1) using nonlinear regression methods to treat the problem of unequally spaced examinations. Furthermore, in many instances this extension will cover the case of missing data, since the presence of missing data generally results in unequally spaced examinations even if the examinations are equally spaced by design.

Relatively little data exist concerning the effects of smoking during adolescence on lung function. However, all such studies[13-18] have suggested that cigarette smoking during this period of life has observable effects. Although none of these studies has observed changes in FEV_1 of the magnitude predicted by the model proposed herein, comparison of results is difficult, since most of these studies[14-18] have been cross-sectional in nature and have depended upon retrospective assessment of the lifetime smoking histories of their subjects. Nevertheless, the results in the example are consistent with published data in the suggestion that the effects of cigarette smoking can be manifest after relatively short periods of exposure early in life.

ACKNOWLEDGEMENTS

The authors would like to thank Vincent Carey for programming assistance. This research was supported by NHLBI research grant HL22528, SCOR grant HL19170 and NHLBI institutional national research award HL07427. Dr. Weiss is supported by NHLBI clinical investigator award HL00740.

REFERENCES

1. Korn, E. L. and Whittemore, A. S. 'Methods for analyzing panel studies of acute health effects of air pollution', *Biometrics*, 35, 795–802 (1979).
2. Whittemore, A. S. and Korn, E. L. 'Asthma and air pollution in the Los Angeles area', *American Journal of Public Health*, 70, 687–696 (1980).
3. Anderson, T. W. *The Statistical Analysis of Time Series*, Wiley, New York, 1971.
4. Rosner, B. 'The analysis of longitudinal data in epidemiologic studies', *Journal of Chronic Diseases*, 32, 163–173 (1979).
5. Korn, E. L. and Whittemore, A. S. 'Methods for analyzing panel studies of acute health effects of air pollution', Stanford University Department of Statistics, Technical Rep. No. 25, 1979.
6. Laird, N. M. and Ware, J. A. 'Random effects models for longitudinal data', *Biometrics*, 38, 963–974 (1982).
7. Afifi, A. A. and Elashoff, R. M. 'Missing observations in multivariate statistics—I. Review of the literature', *Journal of the American Statistical Association*, 61, 595–604 (1966).
8. Anderson, T. W. and Hsiao, C. 'Formulation and estimation of dynamic models using panel data', *Journal of Econometrics*, 18, 47–82 (1982).
9. Rosner, B. 'Multivariate methods in ophthalmology with applications to other paired data situations', *Biometrics*, 40, 1025–1035 (1984).
10. Tager, I. B., Weiss, S. T., Rosner, B. and Speizer, F. E. 'Effect of parental cigarette smoking on pulmonary function in children', *American Journal of Epidemiology*, 110, 15–26 (1979).
11. Tager, I. B., Weiss, S. T., Speizer, F. E. and Rosner, B. 'Longitudinal assessment of the relationship of parents' cigarette smoking and levels of pulmonary function in children', *American Review of Respiratory Diseases*, 125, (4, part II), 145 (1982).
12. Ferris, B. G. 'Epidemiology standardization project', *American Review of Respiratory Diseases*, 118, (6, part II), 55–83 (1978).
13. Tager, I. B., Muñoz, A., Rosner, B., Weiss, S. T., Carey, V. and Speizer, F. E. 'Effect of cigarette smoking on the lung function of children and adolescents', *American Review of Respiratory Diseases*, 131, 752–759 (1985).
14. Backhouse, C. I. 'Peak expiratory flow in youths with varying cigarette smoking habits', *British Medical Journal*, 1, 360–362 (1975).

116

15. Enjeti, S., Hazelwood, B., Permutt, S., Menkes, H. and Tery, P. 'Pulmonary function in young smokers', *American Review of Respiratory Diseases*, **118**, 667–676 (1978).
16. Lim, T. 'Airway obstruction among high school students', *American Review of Respiratory Diseases*, **108**, 985–988 (1973).
17. Peters, J. M. and Ferris, B. G. 'Smoking, pulmonary function and respiratory symptoms in a college age group', *American Review of Respiratory Diseases*, **95**, 774–782 (1967).
18. Seely, J. F., Zuskin, E. and Bouhuys, A. 'Cigarette smoking: objective evidence for lung damage in teenagers', *Science*, **172**, 741–743 (1971).

117