DATA PROCESSING IN THE NATIONAL HEALTH INSURANCE STUDY

David H. Stewart and Martin Seda

February 1978

P-5926

The Rand Paper Series

Papers are issued by The Rand Corporation as a service to its professional staff. Their purpose is to facilitate the exchange of ideas among those who share the author's research interests; Papers are not reports prepared in fulfillment of Rand's contracts or grants. Views expressed in a Paper are the author's own, and are not necessarily shared by Rand or its research sponsors.

The Rand Corporation
Santa Monica, California 90406

DATA PROCESSING IN THE NATIONAL HEALTH INSURANCE STUDY

David H. Stewart* and Martin Seda**

February 1978

*Head, Computer Services Department, Rand Corporation
**Manager of Data Processing, Health Insurance Study, Rand Corporation

ACKNOWLEDGMENTS

We would like to recognize the many contributions of the various members of the HIS data processing staff who, over the last five years, have contributed heavily to the concepts discussed in this paper. Donna Ito, Andrea Schultze and Marilyn LaPrell are deserving of special mention for gracefully coping with the preparation of the manuscript.

CONTENTS

I. INTRODUCTION

This paper was prepared at the request of the Office of the Assistant Secretary for Planning and Evaluation (ASPE) in the Department of Health, Education and Welfare (DHEW) as input to a workshop focusing on the issues and problems in processing data from a longitudinal panel.

Chapter II of the paper reviews the purpose of the Health Insurance Study panel, the evolution of the data collection approach, and the volumes of data being collected. It concludes with a description of the taxonomy of HIS Data Processing.

Chapter III addresses a number of general concerns that the developers of the HIS data processing system feel are relevant to processing the data of any longitudinal study.

Chapter IV discusses particular parameters of design that are reflected in the HIS data processing system. Chapter V reviews the architecture of the HIS system in terms of its component subsystems.

The final Chapter VI discusses several topics of specific interest to the workshop within the context of the established design goal and implementation of the HIS system.

II. OVERVIEW OF HIS

A. EXPERIMENTAL GOALS

Many health insurance programs have been proposed by Congress, labor unions and consumer groups. The approaches taken in these programs vary widely, with some proposing federal insurance plans while others advocate privately administered plans. Regardless of the administration methods, the Federal Government will inevitably play a role in determining the structure of the nation's health insurance plans. Current knowledge of health economics is not sufficient to reliably predict the effects of public policy decisions related to health care financing.

The Health Insurance Study (HIS) is designed to collect, organize and analyze information for developing a fuller understanding of health care economics. To carry out this task, a data bank resource is being created from data gathered in a study involving 2800 families that have been selected and enrolled in different types of of health insurance plans in 6 different geographical locations. Each of these families is having their personal health attitudes and economic decisions monitored on an ongoing basis for 3 to 5 years.

B. PROCESS VIEW OF HIS

During the life of the HIS the methods of data collection have changed significantly (at least from the viewpoint of the data processing needs). The original data collection design was structured around in-person periodic interviews with the participants and receipt of a steady flow of medical claims from the health care providers. The original set of data collection tools consisted primarily of:

- Demographic screeners for eligibility
- Baseline interviews for sample allocation
- Enrollment interviews
- Quarterly interviews in the home
- Medical claims
- Baseline Physician Capacity Surveys

Due to increasing costs, methodological considerations and the expanded interests of the study, a number of alterations to the original concept were made. With the exception of baseline, enrollment and exit interviews, much of the data are collected by mail through self-administered periodic questionnaires rather than in-person interviews. Additionally, extensive self-administered medical history questionnaires are included at enrollment and exit and physical examinations are also performed at these times. To audit the claims reports and collect associated information not available from claims, a self-administered Health Report mailout questionnaire is used on a biweekly basis. The set of data collection tools now includes:

Professionally Administered

Baseline Interviews
Enrollment Interviews
Exit Interviews
Medical Claims
Physical Examinations

Self-Administered

Periodic Employment Questionnaire
Health Care Questionnaire
Annual Income Questionnaire
Knowledge of Coverage Questionnaire
Medical History (Enrollment, Exit)
Health Reports (biweekly)

Miscellaneous Nonparticipant Surveys

Factor Price Index (Phone and mail)
Consumer Price Index
MD Capacity Utilization (Phone)
DDS Capacity Utilization (Phone)
Insurance Abstraction

These changes in the methods of data collection in conjunction with the ongoing revisions to data collection documents have resulted in a significant proliferation of document types. A data processing system designed for a study such as HIS must be able to cope with changes in data collection methods as well as provide mechanisms for structuring, tracking and understanding the effects upon the data base resulting from them.

C. DATA DIMENSIONS

In order to compare the size and complexity of the HIS with other longitudinal studies, the following statistics regarding data dimensions are provided:

1. 11-12 year effort (due to staggering of startup)
2. 6 locations
3. 2050 experimental families and 750 controls: 60 percent of experimental sample participates 5 years; 40 percent 3 years.
4. 2.9 persons per average family
5. 8120 persons in the study by the end of the 1977 fiscal year
6. 46,000 personal interviews performed for the study
7. Average of 650 data elements (net) per interview per person
8. Mailout interviews per family per year:
   - 2 Periodic Employment Reports
   - 1 Annual Income Report
   - 1 Health Care Questionnaire
   - 0.5 Knowledge of Coverage Questionnaire.
9. Data sources and volumes other than interviews include:
   - Medical claims processed - approximately 350,000 for the total sample over the duration of the study.
   - Health Reports processed - approximately 275,000 for the total sample over the duration of the study.
   - Miscellaneous surveys - 20,000 involving the collection of non-participant data over the duration of the study.

10. Approximately 10,000 data elements resulting in 50,000,000 values in final data files.

D. UNITS OF ANALYSIS

In addition to the amount and variety of data observations, the multiple levels and units of analysis is an important consideration in the design of the HIS information system. These include: a) data elements, b) people, c) families, d) households, e) episodes of illness, and f) data collection documents. Since time is a critical experimental parameter, the system has been designed to access data by combinations of the above reference points and accomplishes this with both explicit and implicit reference to time. It is the collecting, indexing, archiving, managing and retrieving data in these many dimensions that constitutes the HIS information science problems.

E. TAXONOMY OF HIS DP

A number of similar studies pertaining to the policy issues of negative income tax have been established [1, 2, 3, 4]. These include the New Jersey, the Rural, and the Seattle/Denver experiments. HIS has a similar pattern of data collection and usage which may be separated into 9 stages: 1) Hypothesis formulation, 2) Questionnaire development, 3) Questionnaire administration, 4) Quality control, 5) Data reduction and coding, 6) Data cleaning, 7) File creation and maintenance, 8) Data extraction, and 9) Data analysis.

In previous experiments, the data processing effort supported many of these stages. In the HIS, however, an information system design has been developed that systematically addresses all of these stages.

Central to this unified approach is the concept of a unified data base utilizing a Data Element Dictionary (DED). As shown in Figure 1, the control point of the HIS data base is the DED. It describes the data base contents, sources, and constraints. It has been developed to improve the organization of information processing and to assist in shortening the learning time necessary for a researcher to become productive with the data base. Many aspects of the DED are discussed throughout this paper, but a brief overview is given here.

Hypothesis formulation defines the data elements to be collected. This stage dictates the basic content of the dictionary. The questionnaire development process describes in detail the location of the data element, its context sensitivity, and the type of value allowed. The DED contains the definition of the item being measured and the mechanism used to measure it. For questionnaire administration, quality control, and data reduction/coding, the DED is a formal reference of proper usage, meaning and manipulation of the data elements.

During data cleaning there is a two-way information path to the DED. As in the previous stages, the dictionary serves as the source for determining data element value correctness. Since perfect data quality is not easily obtained in survey work, the data cleaning process also contributes to the dictionary measures of the varying qualities of data actually obtained.

The file creation and maintenance stage uses the DED to obtain information on data element usage. This stage adds additional quality information to the dictionary and develops statistical profiles on each data element in the data bank.

When the researcher is abstracting a subset of data for detailed analysis, the DED is used to review data definitions and data descriptors.

The analysis effort uses subsets of the data base. Through the data base marginal statistics in the dictionary, the statistical relationship of data elements in the subset to the total data base may be understood.

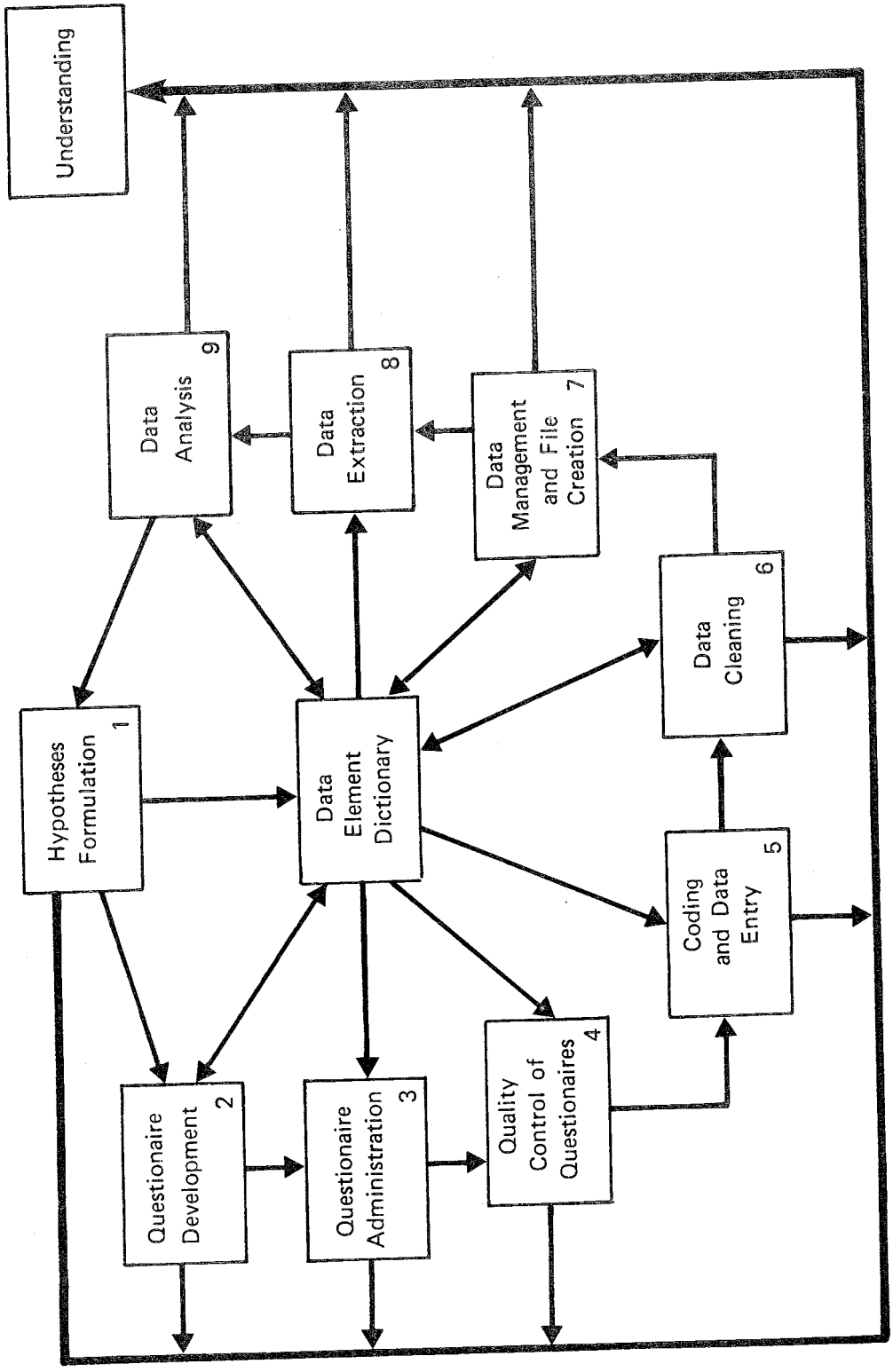

Fig. 1—Data life cycle

III. OUTSTANDING PROBLEMS IN SURVEY DATA PROCESSING

There are a number of problems inherent in the processing of information on projects such as HIS as well as most large survey efforts. These problems preclude complete understanding and total resolution, although they play critical roles in the success of the data processing activities. Those that were carefully considered in the design of HIS data systems are offered here for consideration.

A. DATA BASE ORGANIZATION AND USER COMPREHENSION

With large longitudinal panel-based research projects like HIS, the ability to conceptualize, document, and functionally organize the data base in a manner permitting the principal user of that data base, "the research analyst," to fully comprehend the data resource is critical. Historically, this has not been a problem with which survey researchers were forced to contend. In an era of one-time surveys consisting of a single survey document, the computer files and documentation could be easily constructed to reflect a structure similar to the questionnaire itself.

In contrast, large modern survey efforts tend to be characterized by a great number of document types, different and sometimes conflicting units of analysis, a large number of data elements and conditional logic affecting both the presence of data elements and the presence of entire documents at multiple points in time. Since these projects frequently exist over multiple sites and many years, changes in data collection documents may occur from site to site, and from year to year that were unplanned or are not equally representative over the entire data base.

In a simple survey, a relatively restricted set of hypotheses is being tested and the needs of a small set, if not only one set, of research disciplines are to be satisfied. The dependent and independent variables are more perfectly understood, and normally relate very closely to the measured variables in the survey itself. Frequently, the scope of the entire interview is restricted to the point that the function and eccentricities of the measured variables are within the capacity of the average investigator to retain and recall.

The complexities in projects like the HIS, however, result in a body of data so immense and varied that it is impossible for any research analyst to fully comprehend. There is also interest in attempting to perform analyses which were not explicitly considered in the process of constructing the interviews. Thus, the data base becomes an entity for analysis and there emerge questions about the kinds of hypotheses which the data will be suitable for testing. Over time, the structure of the data base may also become of as much interest to the research analyst as the responses themselves.

For all these reasons, it becomes necessary for the data processing system to provide mechanisms, tools, and systems for organizing and presenting the data to the user beyond those simple techniques that are appropriate for a single survey-document model. The data processing industry terms this kind of support as data base administration. It is necessary that each of these projects begin with an

understanding that a formal data base approach to the management of its data is a necessity.

B. DATA ELEMENT DEFINITION AND TRACKING

Historically, a simple codebook reflecting the single questionnaire and accompanying file was sufficient data documentation. With the complexity of HIS-like data bases the codebook is no longer a sufficient mechanism. Designers of large surveys must appeal to formal data base methods and develop rigorous definitions of data elements as well as formalized data element labeling techniques. Techniques other than positional format for identifying data elements both within their literal context of the questionnaire and within the data base must be developed. This identification can be used to link to the centralized repository for understanding the context of collection and the questionnaire designer's or analyst/user's intention for collecting it.

Mechanisms are required apart from the questionnaires and the data files for tracking the data elements that have been collected. These include: the data element names, the allowable response set, the intended sample space (by this we mean if the data element existed in a conditional logic set, which respondents would be expected to answer the questions), and the locations the data element would be expected to be found in use space—that is, the questionnaires and the questions where it occurs.

C. UNIT OF ANALYSIS DEFINITION AND TRACKING

In complex analytical undertakings such as HIS, there are analysts from different disciplines concurrently using the data base for their research purposes. Frequently, this results in a number of very different and changing units of analysis.

Most statistical/analytic techniques require that the data base be ordered by the unit of analysis. With multiple and changing units of analysis, the data processing system must be able to deal with primary units of analysis explicitly while providing enough flexibility to handle multiple secondary units of analysis. A great deal of the difficulty associated with matching panels in longitudinal studies arises directly out of poorly defined units of analysis and inflexibility in system design.

D. DEALING WITH A GROWING DATA BASE

In longitudinal panel studies the data base available for research analysis is extremely dynamic. Large sections of data are being added to the data base on an ongoing basis. This phenomenon can place a heavy burden on the analyst to know the point in time that he formed his analysis data base, and to interpret the data from that perspective.

Traditionally, analysts have been trained on static data sets with known statistical properties. With a growing data base, a statistical pattern discovered at one point in time may well disappear as the associated data values get added to the data base.

E. STATISTICAL PROCESSING WITH UNCERTAIN DATA QUALITY

Another property of survey data bases and longitudinal panel studies in particular, is uneven or uncertain data quality. Examples of this problem include missing data, refusals to questions, unanticipated responses, and inapplicable questions. Other uncertain qualities can arise from data values that exceed ranges, although the value is verified and correct. These problems are compounded in longitudinal studies where respondents can appear logically inconsistent with themselves in time.

Similar to the problem in the growing data base, researchers are not trained to deal with this phenomenon directly. For example, statistical procedures and protocols have been developed to assist the researcher in treating missing data cases in his analysis. On the other hand, these procedures may differ depending on whether the question is inapplicable or refused rather than randomly missing. Therefore, it is necessary for the researcher to be able to distinguish between various data quality statuses.

Since these problems are frequent in survey data, there is a danger in loss of sample size because of observations containing questionable data. In the HIS information system design, we have provided several features to assist in dealing with the quality problem explicitly. *Every data value has an associated quality indicator allowing the manipulation of data by quality.* Also, the subset formation process allows the researcher to include data quality as a dimension of the subsetting criteria. This means that data values or records can be included or excluded on the basis of variable quality states.

Another important concept is the complement of the subset feature allowing for value substitutions on the basis of the quality indicator. An additional mechanism has been established in the HIS information system that will allow the user to use the quality indicator to specify that a given response should be regarded as questionable without altering its value.

F. PROJECT SKILLS AND TURNOVER OF PERSONNEL

When the projected lifetime of a project like HIS (11-12 years) is considered in the light of the mobility of the American populace, a concern arises. How many key personnel involved in the initiation of the project will still be participating during its projected 12th year of analysis? The history of data processing is littered with noble efforts that were operable only by the computer technicians who built them. Lack of regard toward nonbuilder operation and poor documentation are often cited as primary causal factors.

This same phenomenon also affects the users of information systems. Many times, the information output has only been intelligible to those setting the requirements initially. Given the complexity of the experiment, the design, and the data being collected in the HIS, there is a real danger from shifting of personnel in the analytic ranks as well as in data processing. Factors such as the learning curve associated with using the data base productively are involved in this issue.

IV. DATA PROCESSING DESIGN CONSIDERATIONS

There are a number of key concepts characterizing the design of the HIS information processing system. These concepts have been drawn from the problems discussed in Section III as well as more technical aspects of system design.

A. EVOLUTIONARY SYSTEMS

With an effort as long and dynamic in scope as the HIS, it is naive to believe that the information processing systems for the life of the project can be designed and implemented during the initial stages. At issue is the ability to easily modify various subsystems in response to shifting needs.

The adaptability of an information system must be a primary design consideration and depends heavily upon the tools and techniques used in implementation. Among the common facets of such an evolutionary design are straightforward interfaces and modularity of processors.

B. USE OF DATA PROCESSING INDUSTRY STANDARDS

In the process of implementing and operating the data processing systems for HIS, tools and techniques employed are clearly in the "mainstream" of the computing industry. A goal was to avoid the development of "orphan systems" (i.e., systems operating solely on unpopular computer systems or implemented with programming techniques not universally available). Use of data processing standards implies using techniques and tools with clear, widespread recognition and acceptance. These standards also extended to the manner in which the systems were implemented, tested and documented.

C. CONFIDENTIALITY

The process of collecting micro data from families and individuals on matters of income, health and social attitudes over extended periods of time (3-5 years) yields a dossier on every experiment participant. Rights to privacy demand that the anonymity of research participants be preserved. All information processing systems implemented and operated by the HIS are guided by the *Fair Information Practices Codes* established by the HEW committee on privacy [5]. The types of information on families and persons is categorized into classes of sensitivities and each information process is identified by the classes of information processed. A new operational mode was developed for both computer and manual processing of sensitive information. The goal is to make the careful handling of information so routine that it becomes second nature.

D. SURVEY INSTRUMENT DEVELOPMENT AND DATA PROCESSING INTERFACE

Data processing begins with the measurement instrument or questionnaire. If the organization of the questionnaire precludes satisfactory processing then the analysis effort is lost before it is begun.

Survey instruments are a representation of the data space of interest to the researcher. There are problems of organizing and processing these data so that desired context sensitivities are maintained.

In the organization of complex measurement instruments, such as those being developed in the HIS, it is imperative to clearly identify responses with individuals and to organize skip logic patterns so that the interview does not contain unwanted question paths, thus losing the context sensitivity desired.

The HIS data processing staff has taken an active role in the review and critique of draft survey instruments, and is using the Data Element Dictionary to maintain a history of individual questions and data elements.

E. QUERY CATALOGING

One of the tools developed for data analysis in the HIS saves the selection criteria used by the researchers in extracting information from the data base. This has a twofold purpose. First, it permits the request to be rerun against the accumulating data, allowing the analyst to perform ongoing analysis without reformulating the request each time. Second, it allows the analysts to be cognizant of the work of others, thereby avoiding duplication of efforts. The file of requests is available to all the analysts, and reports are issued to keep the analysts abreast of current subsets being created.

F. GENERALIZED RECORD

A record structure for the data files has been developed permitting data from all data collection documents to be encoded into a single self-defining record format (see Figure 2). Its primary components are a set of record keys or indexes describing the person, family, document, etc., and a list of unique data element identifiers that define the data in the record. Associated with each data element identifier is a data quality indicator. The presence of data element identifiers signals either that the data element value is contained within the record, or that according to the logic of the interview, it should have been. The data status explains the quality of the existing value or the reason for its nonexistence (i.e., refusal, missing, don't know, etc.) In this manner, data compression is achieved as well as record structure-data collection document independence. The advantage of this approach is that all files are structured identically, permitting utilization of a single set of retrieval software operating at the data element identifier level. This also facilitates significant restructuring of the data base without changes in software.

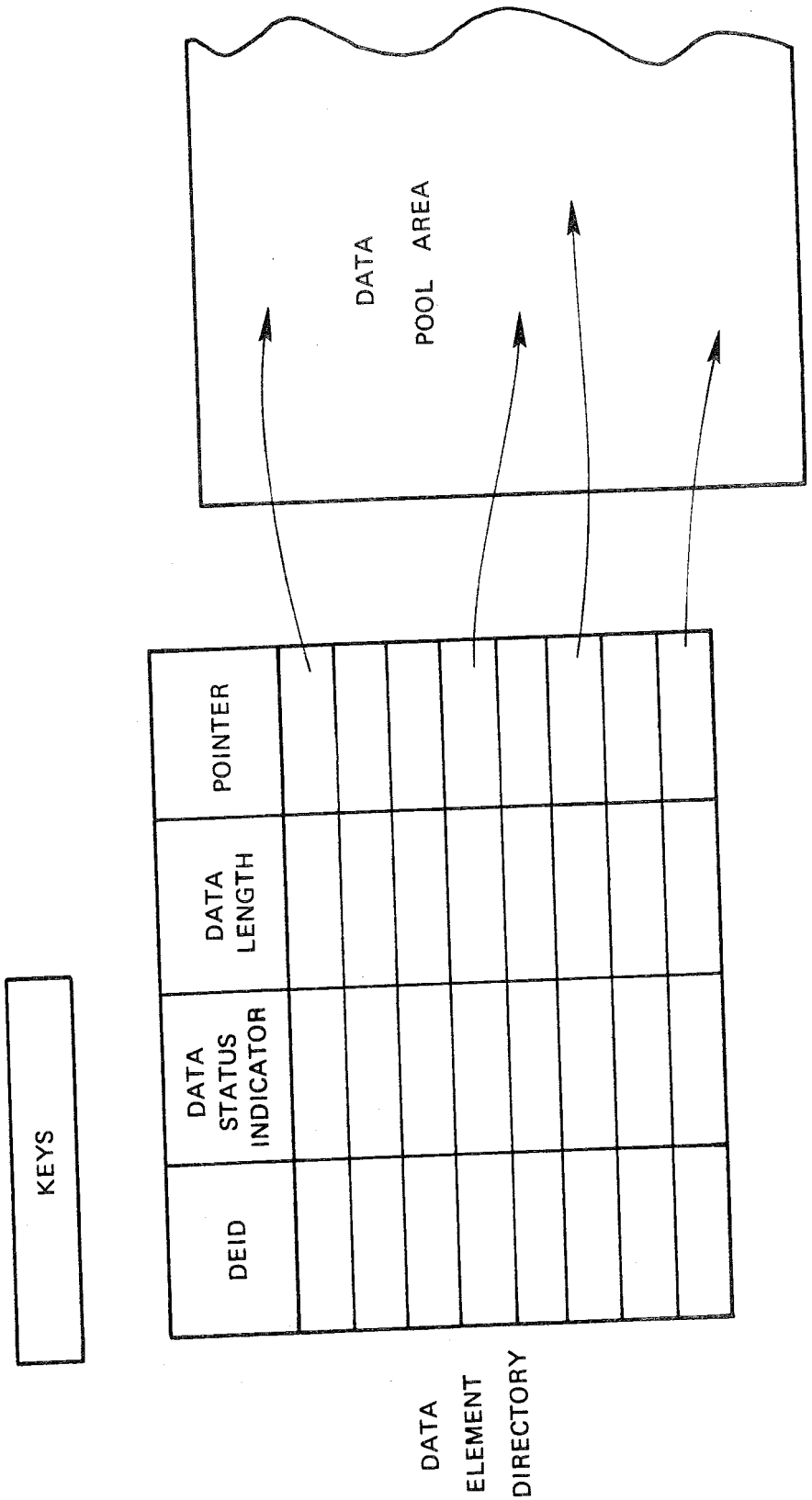

Fig. 2—Generalized record structure

G. DATA TRANSFER

All data are transferred from the field and survey operations to Rand through a single file: the Rand Master Input Tape (RMIT). To ensure that transfer is orderly and to provide for data accountability, an information transfer system was designed and implemented. It has three primary components: 1) program modules at field and survey operations to restructure data from fixed rectangular files to the HIS information structure; 2) an interface tape file (the RMIT); 3) a system at Rand to perform final verification of the data items.

The restructuring module is used to retrieve items of information from the field and survey file and restructure them so they are organized by persons and families, utilizing the generalized record structure.

H. DATA ELEMENT TRACKING

Since the HIS is collecting a large and varied amount of data, it is important to not lose track of how and when information is collected. To accomplish this, a means of data element tracking was developed. For each data item collected during the course of the study, a history is kept indicating:

a) the document(s) on which the data element appears
b) the date of the above document(s)
c) the survey question which was asked to elicit the data element response
d) the allowable responses for the data element on the document

With the above information, the system can provide historical perspective to an analyst or interview designer.

I. MULTIPLE EDITS OF DATA

There is consensus among many survey data users that cleanliness of the data is a noteworthy problem. With this in mind, the HIS system design specifies three levels of quality editing of the data.

The first level of editing, called syntactical editing, reviews data records to assure that mandatory data elements are present and that skip logic chains are followed. This edit also guarantees that numeric answers are numeric and that values are within expected ranges.

The second level of editing concerns the internal consistency of data values. This form of editing, called a semantic edit, checks internal consistencies of data values. Examples include summing the incomes of all working family members and comparing the result to the stated family income.

The third level of editing to be employed is called statistical editing. It isolates those data value groups that are acceptable from the syntactical and semantic edits but are statistical outliers. This form of editing is used to compare data groups against expected values and look for systematic bias resulting from problems of questionnaires or interviewing techniques.

J. QUALITY FLAGS

Another concept that has been developed for the HIS is a quality indicator associated with every data value collected. This indicator gives information about the absence of data or the quality of the data when present. It allows the analyst to subset experimental groups in terms of missing values as well as by a number of quality levels pertaining to the thoroughness with which particular values have been checked. A file containing a statistical summary of data qualities encountered for each data element collected is maintained and provides the analyst with a valuable means of estimating the usefulness of particular data elements.

K. HOUSEHOLD/FAMILY, AND PERSON IDENTIFIERS

Since the HIS is a panel study, the need for a process to identify and track persons and families through the course of the study is mandatory. Additionally, more than one family can exist within a single household and a procedure is necessary to link them together. Without a means of unique identification, flexible data linking between and within these units of analysis would be virtually impossible. Toward this end a method was developed for assigning identifying numbers to households, families and persons in the study. The procedure assigns household, family and person numbers to each individual. The person number for an individual remains the same throughout the study, but his household and family numbers may change as a result of changes in the household/family unit. All three numbers are assigned by computer, and neither one contains any identifying information about the individual or the household/family unit composition.

L. HOUSEHOLD/FAMILY, AND PERSON TRACKING

In the Health Insurance Study, the principal unit of analysis is the person within the context of the household/family decision unit. A major problem in effectively analyzing micro-data collected as a result of panel studies is that a household/family decision unit, interviewed at one point in time, may change from one survey to another. This point has been noted by Harold W. Watts (4) who states the problem as follows:

> "The full complexity of such data is reached when the information is collected for a series of points in time. A person is indivisible for these purposes; he is however, born and he does die. And, the decision units of which he is a part can also change from one survey to another. This process—of birth, death, and mutation of multi-person units—is what makes it difficult to organize, store, and work with micro-data. Different analyses are likely to apply to different decision units or even different versions of what is nominally the same unit."

To manage this problem, the Unit Tracking System has been developed that provides a automated mechanism for tracking household/family composition changes and associating persons to several occurances and types of household/family decision units through time.

V. OVERVIEW OF SYSTEM ARCHITECTURE

The HIS information system is composed of 7 major subsystems, responsible for processing the experimental data at various points between data inception and analysis. Each subsystem is designed to access a number of data files local or unique to its processes and interface to other subsystems through mutually common files. In this manner, each subsystem is relatively independent of all others in terms of computer program interdependence. This design toward flexibility and evolution makes it possible to substitute programs and procedures for each subsystem, with only minor effects upon the organization or implementation methods of other subsystems. Figure 3 shows the file interdependence between these subsystems.

The 7 major subsystems are:

- Data Collection Support
- Information Transfer
- Archival Services
- Data Base Management
- Data Retrieval and Abstraction
- Data Analysis
- Unit of Analysis Tracking and Management

The data base is comprised of 7 main file types. These are:

- Experimental data
- Unit tracking data
- Data Element Dictionary
- Data Element Statistics
- Unit of Analysis Tracking data
- Working Data Bases
- Work files

Following is a brief overview of each subsystem and a description of its relationship to other subsystems. Each subsystem's use of the relevant major file types is also discussed.

A. DATA COLLECTION SUPPORT

The data collection support system is responsible for all aspects of the collection of data from survey procedures as well as the accounting functions related to claim processing. Among the functions performed by this subsystem are:

1. Maintenance of participant name-address lists.
2. Computerized prerecording of interviews.
3. Computerized reports concerning mailed self-administered questions.
4. Questionnaire tracking.
5. Data reduction and cleaning.

6. Claims processing.
7. Generation of unique person and family numbers.

There are two principal subsystems comprising the data collection system. In the area of survey data reduction and cleaning, an extended version of the RECODE system developed by National Opinion Research Corporation is being used. The other functions of the data collection support system are performed by custom programs prepared by the principal field subcontractor who maintains sample lists and processes and pays medical claims.

B. INFORMATION TRANSFER SUBSYSTEMS

Rand receives input tapes from the field and survey operations in the form of the Rand Master Input Tape (RMIT). The RMIT is structured in the self-defining record format. These tapes are processed by the Rand-developed Tape Verification System (TVS). The primary function of the TVS is to identify errors in the data before updating the data base.

The RMIT is constructed from the responses given on the data collection documents. Documents can apply to an individual respondent or to a group of respondents (hereafter referred to as a responding unit). When processed, the responses are divided into separate records for each individual along with a record containing the responses of the responding unit as a whole. The data content of each record is dynamic, being dependent on the questions a respondent is asked and the responses to those questions.

As a result of this dynamic content, each record embodies a logical structure based on key question responses. In addition to the structure within a record, there is a structure through the entire set of records for a responding unit. To ensure cleanliness of the data, it is necessary to check the attributes of individual data elements, and also to check the logical structure of the record and its relationship to the other records for the responding unit.

During execution of the TVS, records are processed by document type within responding units. When errors are located within any of the records from one responding unit for a document, the entire set of data elements for the document are prevented from being entered into the data base until the errors are corrected. The TVS provides a listing of all records processed with identification of any errors which have occurred within each record and responding unit.

C. ARCHIVAL SERVICES

The systematized archive is one of the more important new directions in this system. This importance is reflected in the basic system architecture as shown in Figure 3. A brief review of some of the functions performed will illustrate its architectural importance.

The processing programs, Data Element Dictionary (DED) and Data Element Statistics File (DESF) serve as a control point for:

- Data element definition and naming.
- Locating data elements in all documents.

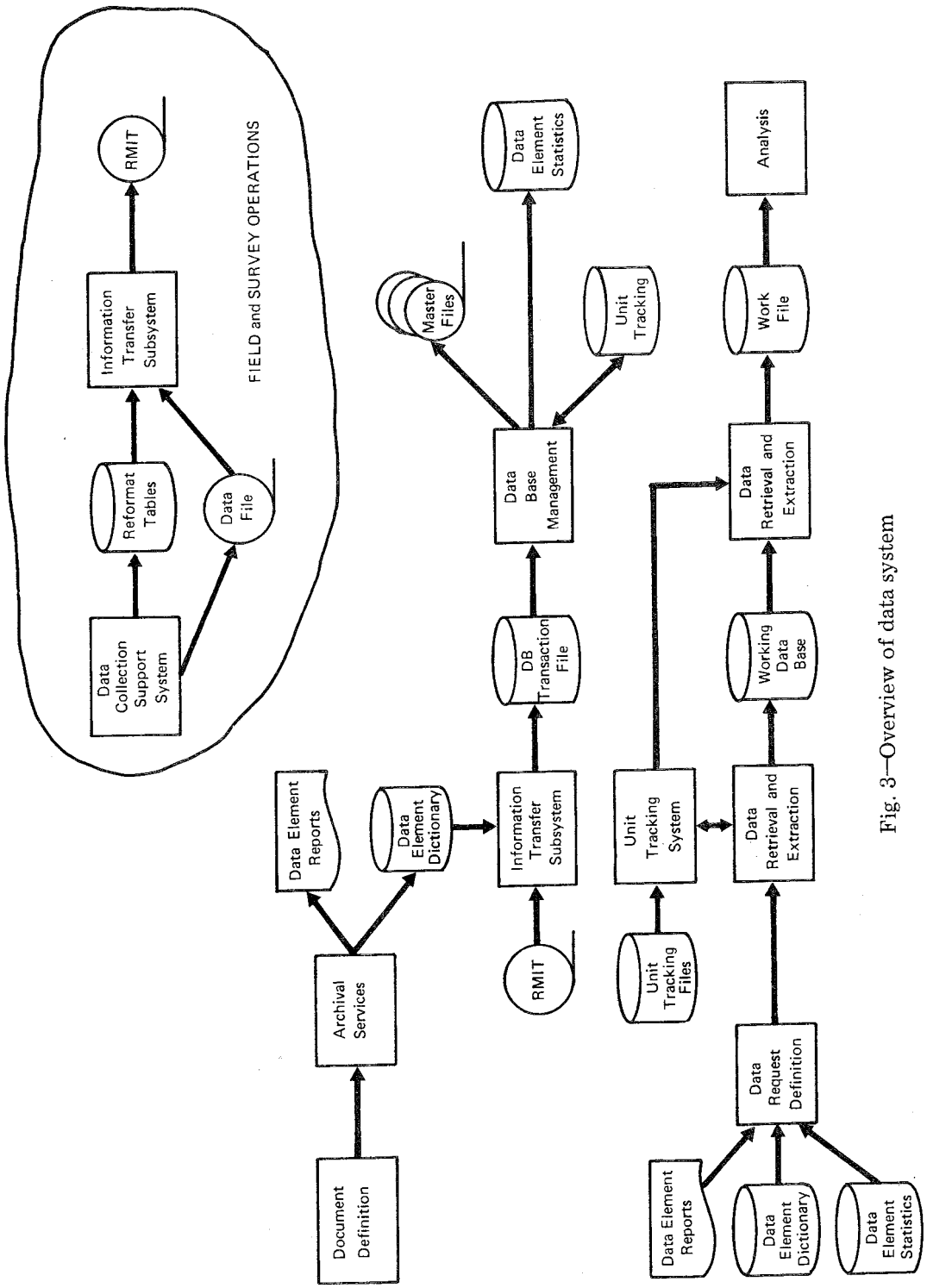

Fig. 3—Overview of data system

- Maintaining a computerized shelf listing of all data collected in the study.
- Describing the editing controls and quality indicators applied to the data.
- Recording and tracking inter-data element relationships.
- Presenting a first level view of the data contained in the data bank.
- Reviewing retrieval requests from the data base.

The archival subsystem is built around two closely associated files—the DED and the DESF. The DED contains detailed information on all document responses (called data elements) gathered during the course of the experiment. For each data element, the dictionary contains information such as:

- Data element length.
- Allowable value range.
- Questions used to gather data.
- Allowable answers.
- Documents containing the data element.

The DED, which is maintained by the Data Element Dictionary Update System (DEDUS), specifies the constraints and characteristics of the data for the Tape Verification System in confirming the accuracy of the data, and contains the archive of descriptive information.

The input to DEDUS is the description of a document as a set of information requests designed to obtain specific data. Each response to an information request is considered to be an individual piece of data and is identified as a data element. The data elements making up a document are defined on two planes: individually as separate responses and collectively as a document.

Collectively the data elements for a document are defined in terms of the logical and arithmetical relationships existing on the document between the data elements. Because a document is administered as a unit, and not as a set of individual, unrelated data elements, the inter-relationships of data elements must be captured. Since the response to one information request can remove the necessity of gathering responses to other requests (skip patterns), these inter-relationships must be properly satisfied to insure the accuracy of the data. When these relationships are defined to DEDUS, it generates logical and arithmetical checks which are stored in the verification system to be used in determining the accuracy of any data received from that document.

The DESF is updated and maintained by the data base update module. It computes the marginals and the quality indicator frequencies found during an update cycle and posts them to the DESF.

D. DATA BASE MANAGEMENT

The data management function is central to the purposes of the data processing effort. In some previous studies, a custom data management system was designed and implemented for the express purpose of the project. HIS has employed general purpose commercially available systems for this purpose 6. The current system in use is Mark IV.

Unlike some other systems developed, the primary activities of data verification (syntactical checks) and validation (semantic checks) are maintained separate-

ly from the data management function. As may be seen in Figure 3, the transaction file is the interface between the information transfer subsystem and the data management system. This transaction file contains only those records of data that TVS has found to be internally consistent and whose data elements are within the correct range in value. The transaction file remains in the generalized record structure of the RMIT.

The data management system uses the transaction file as input, and causes the update of the data bank with the new data records accomplishing the physical movement of entire data records from the transaction file to the master files of the data base.

The design for the data base recognizes the need for the data management function to be evolutionary in nature. The requirements of the data management system may be expected to evolve considerably throughout the life of HIS. The design does not attempt to anticipate the details of all of these requirements in the development of special software and the organization of files. Thus, it is expected that the data base will be reorganized throughout the period of the study and generalized software has been sought to provide this flexibility.

E. DATA RETRIEVAL AND ABSTRACTION

The function of the data retrieval and abstraction subsystem is to provide a mechanism for selecting and retrieving subsets of data from the data base for input to statistical analysis programs and methods. Depending upon the analyst's approach, and the research topic, the data base will be approached in a variety of ways. For example, one analyst may be studying the response to one fielding of a questionnaire, another may be reviewing family behavior over several questionnaires, and another studying a particular group of individuals over the lifetime of the experiment. The data used in each of these different studies is organized into units of analysis by very different criteria. There is evidence suggesting that previous efforts to develop similar social science information systems have found it difficult to provide adequate flexibility for the varying units of analysis and data item interrelatedness required by the analyst. With this in mind, considerable attention has been paid to the development (through interaction with the analysts) of a set of general information structures reflecting the specific needs of the experiment.

Due to the large and diverse nature of the HIS data base, and the different research areas of HIS, the importance of establishing standardized lines of communication between the users and the data base is critical. Each data request involves substantial interpretive costs which may be duplicated unnecessarily without an organized form of communication. To alleviate these problems, a set of protocols have been established to assist the user in requesting data from the HIS data base through a central contact point.

The procedure for requesting data from the HIS data base involves an interaction between the analyst requesting the data and the Data Request Administration (DRA) group of HIS data processing.

The following steps characterize this interaction between users and DRA in extracting data for various analyses:

1. Reviewing the Data Base Descriptive Information

Special information tools which summarize what is available in the data base have been developed to assist the user. These include:

- **Data Availability List.** Contains the processing status of each document, the number of document occurrences in the data file and the file's estimated availability date. It is updated regularly and distributed to all users.
- **Interview and Document Descriptions.** A short description of all interviews and each document used in the interviews. Descriptions include a paragraph of background information, a list of the documents used in the interview, a specification of the sample to which it was administered, and a module specific description of the data collected.
- **Logic Diagrams.** Often an interview involves the administration of several interrelated documents. Logic diagrams are available that describe these relationships in detail showing how they are reflected in the data base.
- **Data Element Dictionary Report.** A report which contains a description and location of all data elements defined in the data base is produced regularly.
- **Descriptive Statistics.** A report which contains marginal statistics for each data element by document on all the data currently in the data base is provided.
- **Work File List.** A description of all the data sets which have been created for users is maintained to aid HIS communication and encourage use of already existing data sets.

2. Initiating the Request for Data

The next step in requesting data is to make a formal request for each set of data desired. A special form is used to provide an overview of the request by describing the type of analysis being performed and basic data requirements relating to that analysis.

3. Reviewing the Request and Preparing an Estimate of Costs

The request is reviewed and ambiguities and problems are discussed. Review is complete when the user and DRA share a common perception of the task. DRA then estimates the cost of the request in terms of data processing days and computer resources and reviews it with HIS management for approval and priority assignment.

4. Requesting a Work File

The requestor documents the data request on a form describing in detail the specifications of the data desired. The details include data elements, combinations of data elements forming created variables, the unit of analysis, and the statistical packages to be used.

5. Reviewing the Final Request and Scheduling the Work

Upon receipt of the work file definition, DRA checks for specification errors and attempts to identify any omissions. If significant problems are found, the form is returned to the requestor to be reworked. When the request has been finalized, DRA reevaluates the cost and time estimates, schedules the request, and the Work File creation effort commences. DRA returns to the user a copy of the request with cost and schedule estimates.

6. Executing the Request

The DRA organizes the request processing tasks for the data base programmers. The DRA specifies the data files to be used and determines the organization of units of analysis using the unit tracking files. Entire records reflecting the entities and time points of interest are extracted from the data files forming a Working Data Base. It is with this Working Data Base that variable value restrictions and derived variable computations are performed. This results in greater overall efficiency in that as Work Files are formed from Working Data Bases, analysts may ask for new or expanded work files without causing the re-retrieval of the records of interest from the master files.

7. Describing the Work File

Upon completion of the file creation, DRA submits a copy of the file documentation to the requestor. This includes a printout of the first few observations in the Work File, marginal statistics, the format of the data, and the physical location of the data set. The Work File format will be processable by the statistical packages specified in the request form.

The file remains a part of the HIS data base and is stored under an HIS system account as a Work File. This data set is maintained and backed up by DRA and a copy of the complete documentation is available for project-wide use.

F. DATA ANALYSIS

Generally, the Work Files are created in SPSS system format. This is a fully self-documenting file containing a directory of the variable names to be used by the analyst, the data element or elements used to form it, a link to the forming algorithm in the data request and the label names to be used on all statistical print out.

Currently, HIS has adopted SPSS as the standard statistical package for data descripton, summary statistics, etc. All further variable or data element combinations which the analysts may form are captured as a result of standard SPSS processing. Thus complete documentation is formed automatically in the analysis process linking each variable to the initial data request and Working Data Base.

SPSS is not sufficient, however, for much of the analysis being performed. When other statistical packages or specially written programs are used, the data for these are extracted from the SPSS file by extraction procedures inherent in SPSS.

In addition to providing the documentation discussed above, this approach provides the added feature of standardizing formats and conventions for display of common statistics. Thus the analyst can easily move among analysis efforts without having to familiarize himself with the reporting mechanisms of each analysis group.

Finally, by using the data file format for analysis of a commonly used statistical system, the exchange of files with researchers at other institutions is greatly facilitated.

G. UNIT OF ANALYSIS TRACKING AND MANAGEMENT

In order to establish (and maintain) the longitudinal property of the data collected by the HIS and assure its usefulness in fulfilling the analytical goals of the study, it is necessary to be able to unambiguously characterize the decision units (or units of analysis in HIS parlance) on which the data has been collected. To facilitate this a Unit Tracking System (UTS) is used.

The purpose of the UTS is to provide an automated mechanism for tracking decision units (i.e., households, families), and the changes which occur to them over time. For the purpose of the HIS, the computerized files established and maintained by the UTS respond to the following needs (each to be described individually below):

- Provide the mechanism necessary for tracking individuals through time.
- Provide the mechanism for establishing and describing the different decision units for the different analyses to be performed on the data.
- Serve as the interface between the analytical and operational facets of the HIS.
- Serve as an index into the analytical data files.

1. Tracking Individuals Through Time

A major contribution to the success or failure of longitudinal studies such as the HIS is the ability to identify, and locate, for reinterviewing purposes, those individuals enrolled in the study. Explicit account must be taken of persons who move in and out of decision units. Subsequent data analyses must have a source for establishing the size and structure of the unit in which the data was collected on these individuals in order for them to be meaningful. It is quite likely that different analyses apply to individuals who change decision units over time and those who do not. Thus, the UTS provides the facility for tracking individuals enrolled in the study both for purposes of reinterview and also to ensure that, regardless of whether or not a person changes decision units over time, the data collected for the person can be readily associated with that person.

2. Describing Different Decision Units

At a single point in time, a person enrolled in the HIS is a member of two decision units (not necessarily disjoint):

- A Family Health Protection Plan (FHPP) unit that includes all family members that are covered by the FHPP policy.

- An economic unit that contains all the family members economically dependent on a designated household head.

The complexity resulting from an individual's membership in two decision units at a single point in time, is that the tracking problem becomes two-dimensional. That is, a person may move from one economic unit to another, while still maintaining membership in the same FHPP unit. It is a function of the UTS to provide the information necessary to establish a person's membership in these decision units, as well as tracking an individual's movement from one decision unit to another.

3. Interface Between Analytical and Operational Facets of HIS

The data files maintained by the UTS are utilized by operations personnel for providing the information necessary to locate persons for both the claims and survey operations of the HIS. For the purposes of claims operations, a person is a member of an FHPP unit; for purposes of surveys, a person is a member of an economic unit. To add to this complexity, different analyses to be performed on the data require that an analyst be able to view a person as being in either an FHPP unit, an economic unit, or both at any point in time.

In order to satisfy both the operational and analytic goals of the study (oftentimes competing goals), it is necessary to allow these units to co-exist without encumbering either the analytic or operational facets of the study. The UTS provides the basis for this co-existence, thereby serving as the interface between the operational and analytic facets of the HIS.

4. Index Into the Analytic Data Files

The general philosophy adopted with respect to the use of the data files associated with the UTS provides two mechanisms by which analysis of the HIS data files can be performed in an atmosphere of constantly changing decision units. The first approach directly utilizes the information contained in the UTS files to separate those decision units which have changed configurations (i.e., added or lost people) from those units which have remained stable throughout the study.

The second approach is primarily interpretive. By providing computerized files containing a complete history of an individual's membership in different decision units, and the time frames in which the membership occurred, the UTS permits the analyst to link together data on persons through time, with respect to their decision units at each point in time.

5. Operation of Unit Tracking

In order to accomplish the tracking function, a well-defined set of rules must be specified indicating the changes that require the system to create a new unit, for both operational and analytical purposes. In order to implement such a function, it is necessary to define the components of the decision unit. For the HIS, a decision unit, whether an FHPP or an economic unit, is comprised of a household which is made up of one or more families, each family containing one or more individuals.

a. Identification Systems. In addition to the set of rules by which change is

defined, a set of conventions were established for identifying the entities of the decision unit. Within the framework of the HIS, each unique household, family and person is assigned unique sets of identifiers. The hierarchical relationship between household, family and person is made apparent through the identifier assignment rules that give all families in a household the same household identifier and all individuals within the family the same family identifier. Uniqueness is further defined as a function of one of the following three identifier systems.

- Participant Series identifiers—the set of identifiers used in field operations and claims.
- Transfer Series indentifiers—the set of identifiers used for transferring data between field and survey operations and Rand. The Transfer Series identifiers function as a confidentiality link file.
- Analytic Series identifiers—the set of identifiers used in the data base.

b. Files. There are three primary files involved in the operation of the unit tracking system. These are: the Person Link File, the Houshold/Family Link File, and the Family Tracking File. Their functions are described below.

- **Person Link File.** The basic purpose of the Person Link File (PL) is to maintain a linking between the transfer identifier and the analytic identifier assigned to each person in the study. The primary function of the file is to provide the analytic identifier for a person when given the transfer identifier. Secondarily this file is used to verify the validity of a transfer person identifier by attempting to link it. Because all valid person identifiers are required to be in the PL, an identifier that cannot be linked is not valid.
- **Household/Family Link File.** The basic purpose of the Household/Family Link File (H/FL) is to maintain a linking between the transfer household and family identifiers which are associated with a particular household/family configuration and the corresponding analytic household and family identifiers. The primary use of the file is to link sets of transfer identifiers into the initial sets of analytic identifiers which are coordinated with them. There are two secondary uses of the H/FL which are made possible because of the nature of the file. By making use of the generic key access method of indexed sequential files, a list of all of the transfer families related to a specific transfer household is obtained. Since all sets of transfer identifiers assigned to households and families in the study are defined to the H/FL, the file is used to verify any sets of transfer identifiers as being valid.
- **Family Tracking File.** The basic purpose of the Family Tracking File (FT) is to relate persons to families and families to households over the life of the study. The FT is cumulative over time; as household and families reconfigure over time, the new configurations are added to the file without removing the old configurations. Each configuration is linked backwards and forwards to both the previous configurations from which it was formed and the new configurations which have been reconfigured from it. There are numerous functions which can be performed by the FT.
  - Because of the linking of configurations through time the FT is used to track the household and family configurations of a person through-

out the study. The household and family relationships of any person at a given point in time can be identified through the FT.

- The tracking capabilities of the FT are used in conjunction with the H/FL and PL files to translate transfer identifiers into contemporary analytic identifiers.
- In reversing the linking process, the FT is used to translate analytic identifiers back into transfer identifiers.
- Because all valid analytic identifiers, and all valid configurations of those identifiers, are contained on the FT, the validity of any analytic identifiers and configurations of analytic identifiers can be verified. Any identifier or configuration which cannot be found on the file is considered invalid.
- Using the generic key access method for indexed sequential files, all the families and persons related to an analytic household identifier are identified.

6. Maintaining the Unit Tracking Files

The files discussed above are updated and maintained by a set of procedures and processes that are initiated by the field and survey operations. They maintain a sample maintenance file that contains participants names, addresses, relationships, etc. This file is organized by participant identifier. Changes in the household/family configuration are detected during survey, claims or other field operations. These changes are evaluated using the formal rules of unit definition. If the changes are of a nature to require establishing either new participant units or analytic units, a transaction is prepared by the field contractors to update the unit tracking systems of the field contractors and Rand. These transactions are processed by computer programs containing the appropriate decision rules, the files are updated, and new household and family members are established wherever necessary.

VI. OPERATIONAL EXPERIENCES

There are several topics of interest to the SIPP Workshop that are not directly covered in the foregoing description and discussion of the HIS data processing system. In conclusion to this paper, some observations of HIS experiences in these topics are made.

A. TIME, MANPOWER, AND EQUIPMENT REQUIREMENTS FOR DEVELOPMENT OF THE DATA PROCESSING SYSTEM

The initial design considerations for the system began in the 1971-72 period. Since that time approximately $750,000, in personnel costs, have been invested in design and implementation of the system. Since additional design and implementation are expected to continue throughout the life of the study, this figure has a possibility of being slightly misleading. In addition, approximately $750,000 in computer time has been invested in the development and implementation of the system. These figures include the programming support developed specifically for the HIS by several subcontractors over the last five to six years.

The system has been developed on mainstream commercial IBM processors of various sizes. Currently, a portion of the survey and field operations is accomplished on an IBM 370/135 and the body of the system discussed in this paper has been implemented on an IBM 370/158 at Rand. To date, the machine utilization of the latter processor by HIS is estimated at approximately ten percent of the system's capacity.

B. LENGTH OF TIME FROM INTERVIEW DATE TO FINAL TAPE

The primary focus of this topic is: What is the minimum time loss from interview date to the delivery of the final tape, given that this is a fundamental objective? Although timely delivery of data for analysis has been an issue on the HIS, greater emphasis has been placed on other considerations. Of all the factors that have affected the scheduling of surveys and the processing of data, minimization of the time between fielding and and the preparation of the clean data file has held only modest priority. There have been, in contrast, extraordinary demands placed upon data processing systems in order to initialize operations at experimental sites and prepare minimum amounts of data so that experimental participants could be enrolled and released from the study. Thus, less of a focus has been placed upon the speed with which an entire interview can be processed. The model used in HIS to measure this timeliness has been "What is the length of time from the last interview date to the final tape for the interview wave?" This takes into account such factors as problem cases which may not be completed until late in the survey cycle, as well as data cleaning philosophies that favor uninterrupted cleaning of the entire interview sample. Using this model it is estimated that the turnaround could be as short as three months. In actual practice, without a high priority, the best turnaround experienced has been approximately six months.

C. ASSESSMENT OF THE PROCEDURES USED TO ASSURE DATA QUALITY: EDITING AND CODING TECHNIQUES

The HIS data goes through two cleaning cycles. The first cycle takes place in field, claim, or survey operations and the final cleaning occurs before the data are entered into the data base. The first phase reviews data for type and value range conflicts, and in the case of questionnaires, for the basic skip logic pattern. The second phase of cleaning (as discussed in chapter V.B.) performs a more extensive form of intra-record consistency checking as well as relating individuals within families. This phase of cleaning consistently turns up problems that the first phase was not able to detect. In panel survey studies one of the most important aspects of quality control is to insure that the respondent is bound correctly to his responses within the questionnaire. Complex physical organization of the data collection documents, such as multiple grids and skip logic between these grids, offer many opportunities to lose control of respondent's identity. The HIS system appears to have provided a more than adequate facility for this type of quality control.

However, the HIS system presupposes that data extraction from documents, as well as initial cleaning and reorganization of the data into the generalized record format, have occurred. In our experience, there is much improvement required in these tools. The most probable opportunities would come out of a minicomputer-based system designed specifically to assist in questionnaire design, data cleaning specification, and provide some form of interactive data entry cleaning and reorganization of the data into the physical format used in the data base. The tools that have been used for these functions in the HIS are only marginally acceptable and are based upon designs and concepts arising from far simpler survey tasks.

One of the difficulties of multiple groups performing cleaning is insuring an adequate level of correspondence among the cleaning processes. In the HIS system design the Data Element Dictionary is intended to provide this correspondence among all processors of the data. In actual practice, the autonomy and the distance of organizations processing the data have resulted in an uneven control over the process. The primary effect is a duplication of effort, and not necessarily a lessening of data quality. Since the dictionary controls the entrance of data into the data base, the system has full and accurate information concerning the data that is allowable in the data base. It would be much more effective, however, if the cleaning processes were integrated to a higher degree.

In addition to the basic technological issues of data quality and editing, some HIS analysts have expressed the view that extensive data editing has the potential to induce serious systematic biases that are difficult to detect. In recent surveys, greater attention has been paid to keeping respondent answers intact even when they do not conform to anticipated responses. One concern of methodologists has been the lack of a practical understanding of how initial values may have been changed to reflect a set of possibly biased cleaning and editing specifications.

The HIS approach to this has been to utilize the data status indicator associated with data element values to encode the exceptions and suspicions noted during the data cleaning activities, rather than actually reconstructing the data to make it "clean." At issue is the extent to which the ability to provide "clean" data frequently presupposes an unlikely robust understanding of the phenomenon being measured. Frequently, a debate has ensued over whether multivariate relationships can be described and predicted sufficiently to be used in cleaning the data. Usually,

such inquiries result in recognition that a theoretical basis for the cleaning specifications that avoids the possibility of inducing extremely complex and systematic biases into the data does not exist.

These concerns suggest that there is a need for a greater understanding of the objectives of data cleaning among data analysts and measurement theorists. At question is the degree of precision necessary to do the analysis and the extent to which further precision is either economically infeasible or exceeds the limits of understanding of the phenomenon being studied.

A review of the objectives of data cleaning and editing could well produce a taxonomy more clearly identifying those error components that are likely induced by data reduction techniques in contrast to those forms of inconsistencies which lie either within the respondents own understanding, or within the interface between the respondent and the data collection document. It may well be that these latter problem responses due to respondent misunderstanding should enter the data base uncorrected, and become a subject of analysis performed by the measurement theorist to allow the area subspecialist to understand the degree of precision existing in the data base.

D. INTRICACY AND FEASIBILITY OF PERFORMING DATA EDIT CHECKS WHICH ARE LONGITUDINAL IN NATURE

The complexity of multivariate edit checks between survey waves was ultimately determined to be extremely problematic after many discussions over the issues of longitudinal editing were held. Initial designs for the information system included this form of data editing because, from a technologist's point-of-view, it seems an obvious requirement and other efforts had appeared to be successful in this kind of activity. The net value to be returned, the degree of precision to be employed, and the general efficacy of longitudinal data editing was reviewed. The consensus reached among HIS analysts was that the requisite precision in the underlying theory did not exist, nor did statistical techniques that were adequate in detecting unlikely longitudinal patterns without producing statistical and operational paradoxes. The operational implications of confirming from the respondent whether one or more of the values measured over time were incorrect are troublesome since they require respondent recall (which the methodologists frequently suspect as a common source of error.)

In addition, where closed systems of knowledge do not exist regarding the phenomenon being measured, the statistical techniques available generally offer greater opportunities for paradoxical discoveries. For example, one may find that the observation in question is in fact correct, which implies that the apparent consistency in previous observations was incorrect, thus upsetting the statistical basis of consistency. To correct the problem, the respondent would have to exercise greater capabilities of recall than methodologists generally believe exists. If the respondent did possess either records or an extraordinary sense of recall, and was able to introduce a consistency within his own response set, then the retrofitting of the data would require a reestimation of the probability of the consistency across responses for other respondents over the same time periods, which in some cases might literally be the total duration of the study. Because of this type of paradox and the costs of tracking and resolving them, HIS is not performing longitudinal editing.

In these deliberations, it frequently appeared that the analysis was being performed without the data to determine what was a good or bad value set. It is an information scientist's view of HIS-like studies that they should be viewed as data collections to be studied, mapping sensitivities and possible inaccuracies as a function of a research discipline. This would permit the analyst to determine what data are usable and what analytic techniques are possible using statistical judgments and techniques.

E. ISSUE OF BOUNDING INTERVIEWS

The original design of data collection and processing included the feedback of responses from previous survey waves to field and survey operations with each subsequent survey wave. With the exception of family configuration information, very little has been done in this regard. This is the result of essentially three considerations. First, the opertional consideration of having to insure that an accurate and adequate return of data would be possible within a quarterly framework turned out to be unrealistic. Considering the lead time that field and survey operations require as well as the length of the fielding period of the previous interview wave (extended to include the receipt of the problem questionnaires), data cleaning and reduction of the full panel would have to be accomplished in approximately one month. Second, if the haste to return information to the field results in inadequate cleaning of the data, there is opportunity for inducing further systematic bias by causing confusion in the respondent. This can cause the respondent to reconsider his answers based upon what was actually an error induced by the data reduction process. Third, the issue of bounding surveys was removed nearly altogether because of the conversion to self-administered questionnaires. Without professional interviewer involvement, it becomes difficult to understand the extent to which information being returned to the respondent affects their subsequent responses.

To the extent that bounding is another mechanism for assuring quality in the data from the respondent, the HIS approach has been to appeal to a nearly independent information flow on those measures of critical importance where over-reporting or under-reporting could seriously affect the analytic findings. The biweekly health report mailout is a prime example of this form of secondary and independent information flow. It permits independent calibration of the accuracy of the principal utilization measure reflected in the claim data.

REFERENCES

1. Brewster, J. A., "Estimates of Personnel Requirements for a National Program of Income Maintenance," presented at the 41st National Meeting of ORSA, 1972.
2. Dugan, D. J., "The Gary Income Maintenance Experiment: Analytical Objectives," Presented at the 41st National Meeting of ORSA, 1972.
3. Primus, W. E., "Data Collection and Processing Problems Associated with Social Experimentation," Presented at the 41st National Meeting of ORSA, 1972.
4. Watts, H. W., "Microdata: Lessons from the SEO and the Graduated Work Incentive Experiment," *Annals of Econometric and Social Measurement*, Vol I., No. 2, April 1972, pp. 183-192.
5. *Records, Computers and the Rights of Citizens*, Report of the Secretary's Advisory Committee on Automated Personal Data Systems, U.S. Department of H.E.W., July 1973.
6. Yormark, B.; Stewart, D. H., *A Data Management System Evaluation for the Health Insurance Study*, P-5181, Rand Corporation, November 1973.