A RESEARCH DATABASE FOR IMPROVED DATA MANAGEMENT AND ANALYSIS IN LONGITUDINAL STUDIES

ROGER A. BIELEFELD, Ph.D., TOYOKO S YAMASHITA, Ph.D., EDWARD F. KEREKES, EHAT ERCANLI, M.S., AND LYNN T. SINGER, Ph.D.

We developed a research database for a five-year prospective investigation of the medical, social, and developmental correlates of chronic lung disease during the first three years of life. We used the Ingres database management system and the Statist statistical software package. The database includes records containing 1300 variables each, the results of 35 psychological tests, each repeated five times (providing longitudinal data on the child, the parents, and behavioral interactions), both raw and calculated variables, and both missing and deferred values. The four-layer menu-driven user interface incorporates automatic activation of complex functions to handle data verification, missing and deferred values, static and dynamic backup, determination of calculated values, display of database status, reports, bulk data extraction, and statistical analysis.

In recent years there has been growing concern among researchers, institutional administrators, and government officials regarding proper data management procedures [1-7]. Much of the failure to adhere to proper procedures in scientific research can be attributed to the complexity of the systems, both computerized and non-computerized, that are used for data management [8]. Often this complexity manifests itself in the requirement for frequent manual data-manipulation procedures that can lead to error. These procedures include editing data for the purpose of locating and deleting records with missing values, editing data for the purpose of recoding and correcting erroneous values, and restructuring data as needed for processing by specialized statistical software. Manual editing can result in multiple versions of the same items of data as well as new errors.

As we set out to create data management procedures and software for a developmental study of infants, we tried to address the issues that historically have had the potential for introducing error.

Background
Bronchopulmonary dysplasia (BPD), the most prominent severe lung disor-
der of infancy in the United States, results from oxygen administration to premature infants with inadequate lung development for survival without assistance. The number of infants and young children who survive formerly lethal lung diseases such as BPD has increased considerably over the past two decades [9], and a growing body of literature indicates that the experience of early severe chronic lung disease is associated with negative developmental sequelae later in life [10–12]. Infants with BPD frequently have low birth weights, prolonged hospitalization, complex nutritional problems, an increased risk of neurologic problems, and delayed growth and development [10–13].

The U.S. Public Health Service and the National Institutes of Health have awarded grants to Rainbow Babies and Children's Hospital for a longitudinal investigation of the medical, social, and developmental correlates of chronic lung disease during the first three years of life. Infants have been recruited into the study by examination of the medical records of neonatal ICU admissions at three hospitals in the Cleveland area, to find infants with a diagnosis of BPD and very low birth weight (<1500 grams). Healthy full-term controls matched with respect to age, sex, and socioeconomic status were recruited from the same hospitals. To date, approximately 350 infants have been recruited into the study.

Demographic information is collected for each infant and its parents (Fig. 1 and 2), and standardized assessments of the infant's developmental and physical functioning are performed five times during the period from birth until the age of three years. More than 35 tests including 1300 variables are administered at each of the five visits. Thus, longitudinal data spanning a three-year period are collected, accounting for well over 2 million separate data values.

**Design Objectives**

The objective of our work was to design and develop a flexible integrated system for data entry, management, retrieval, reporting, and analysis to support the developmental study described above. Special attention was given to avoiding errors and to the problems associated with missing data and statistical processing. At the core of this system is a database designed to support research objectives, the most important of which is to derive statistical conclusions from large volumes of data. Other objectives were the ability to alter the database schema while preserving the usefulness of the data, and to support blind and double-blind studies, longitudinal data, multiple studies of subjects drawn from a population defined by the database, data entry with validation, and the production of routine reports.

Because we expected new related studies to occur and evolve while the database was in use, its design and implementation had to be flexible enough to accommodate change. To handle the simultaneous development of both a research agenda and the software supporting it, our methods include procedures for maintaining current information about both the stored data and the development of the database and associated utilities.

**System Description**

The main characteristics of the BPD database are the length of its records, the presence of longitudinal data in the form of values from repeated tests, storage of both raw and calculated values, and the presence of both missing and deferred values. These features of the database demand well-organized and specialized storage structures.

**Implementation Platform**

Implementation of the database was originally planned for an 8086-based microcomputer with 16 MB of memory and 300 MB of online disk storage. The system ran the SCO Xenix 2.3.3 operating system and provided simultaneous access to up to 12 users by means of both hardwired terminals and modems. The Ingres database management system (ASK/Ingres Corporation, Alameda, Calif.), Release 5.1, was initially chosen as the software base for implementation of the database. The Ingres system is relational [14], and it provides structured query language (SQL) for both interactive (ad hoc) and "scripted" (routine) queries. A Query-by-Forms (QBF) facility allows the user to create custom data entry screens incorporating data validation, and an Application-by-Forms (ABF) facility allows the user to create customized, forms-based applications for queries, updates, and reports. For complicated queries run on a regular basis, Ingres provides a library of SQL routines that can be called from within C programs, and a preprocessor, ESQL, for compiling these programs. Finally, Ingres provides a variety of data structures and access methods for data storage and allows these to be altered without complete recreation of the database. After some initial experience with the database, the storage structures used for the relations were modified to B-trees [15] with unique keys. B-trees were chosen to allow rapid access to records on key fields, acceptable overhead associated with insertion of new records during data entry, and automatic prevention of the entry of records with duplicate keys.

After the initial implementation and while the database was in use, we transported it to an 80486-based microcomputer that had a similar hardware configuration but was running SCO Unix 3.2.4 and Release 6.4 of Ingres. The experience was not overly difficult, largely because conversion utilities had been supplied with the software. One such utility converted each relation and each QBF into the new format. Each frame in the ABF application had to be rewritten, but the code-generation software in the newer release facilitated this greatly.

**Missing and Deferred Values**

We expected that values tied to a particular entity in the database would often not be available for entry into the database at the same time. A distinction would have to be made between "missing" data, whose values are unknown and will never be known, and "deferred" data, whose values are currently unknown but are expected to be known in time for inclusion in future data analysis. Clearly, a deferred value may be reclassified as missing if appropriate. Likewise, a missing value may
### THE PERMANENT DEMOGRAPHIC ENTRY FORM:

<table>
<thead>
<tr>
<th>ID:</th>
<th>HOSP:</th>
<th>(1 Fairview; 2 Metro; 3 RB&amp;C; 4 Other)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GROUP</td>
<td>(1 BPD; 2 LBW; 3 Term)</td>
<td>ELIG</td>
</tr>
<tr>
<td>COCAINE?</td>
<td>(1 Cocaine; 2 Cocaine control; 0 Other)</td>
<td></td>
</tr>
<tr>
<td>AGE ENROLLED</td>
<td>(1 T1; 2 T2; 3 T3; 4 T4; 5 T5)</td>
<td></td>
</tr>
<tr>
<td>Mult. Birth Status</td>
<td>(1 Single; 2 Twin; 3 Triplet; 4 Quad; 5 Quint; 6 oth)</td>
<td></td>
</tr>
<tr>
<td>Birth Order</td>
<td></td>
<td>Sibling Order</td>
</tr>
<tr>
<td>LAST_NAME</td>
<td></td>
<td>Aka</td>
</tr>
<tr>
<td>CHILD</td>
<td></td>
<td>HOSP_NO</td>
</tr>
<tr>
<td>BIO MOTHER</td>
<td></td>
<td>BIO FATHER</td>
</tr>
<tr>
<td>Bdate</td>
<td></td>
<td>Gest. Age (weeks):</td>
</tr>
<tr>
<td>B_Weigh (gr.)</td>
<td></td>
<td>Sex</td>
</tr>
<tr>
<td>Race</td>
<td>SES(birth)</td>
<td>SDS(birth)</td>
</tr>
<tr>
<td>M_bdate</td>
<td></td>
<td>F_bdate</td>
</tr>
<tr>
<td>M_Age_b</td>
<td></td>
<td>F_Age_b</td>
</tr>
<tr>
<td>M_Race</td>
<td></td>
<td>F_Race</td>
</tr>
<tr>
<td>Match1:</td>
<td>Match1 Group:</td>
<td>Match1 ID:</td>
</tr>
<tr>
<td>Match2:</td>
<td>Match2 Group:</td>
<td>Match2 ID:</td>
</tr>
</tbody>
</table>

---

**Figure 1. A screen used for entry of demographic data.**

show up unexpectedly and be reclassified as known. At the outset of our work, we decided to use special “missing” and “deferred” values, outside the domain of legitimate data values for the attribute, to take the place of missing and deferred data in the database. Later, after determination of the schema of the BPD database, we determined that deferred data values did not occur except for groups of values that made up entire records in database relations. Therefore, no special value for “deferred” was required.

We created a “data status” relation to keep track of the entry status of data for each patient represented in the database. In this relation we recorded the absence of data, whether missing or deferred, on a record-by-record basis for all relevant relations. By examining this relation, one can see what data were entered, not entered, or missing at a given time.

**Calculated Values**
The BPD database includes many calculated values that need to be computed by the software automatically. For example, summarized psychological test scores must be computed from raw scores, the age of an infant at the time of a particular test must be computed from the date of the test and the infant’s date of birth, specialized derivations must be computed from raw test scores, and the corrected ages of infants must be computed. Corrected age is of great importance in developmental studies involving premature infants, because elapsed time since conception (rather than chronological age) must be
used as the basis for comparison of development against age. In the BPD database, the gestational age at birth is stored along with the actual date of birth, and the corrected age of a premature infant is defined as the chronological age minus the degree of prematurity of the infant.

We found that the Ingres database management system provided the functionality required for incorporating the automatically updated calculated values. Although storage of calculated values introduces redundancy in the data, we found that the advantages of having such values computed only once rather than each time they were needed outweighed this disadvantage. In addition, the danger of inconsistent values resulting from redundancy is mitigated because the redundant values are calculated, not entered directly or independently. Calculated values are designated as “read-only” database values so that they cannot be changed.

Quality Control
Each variable in the database has a domain of possible values from which its value at any given time must be drawn. This domain of possible values, including any special values that may be used to represent missing and deferred items, must be known to the data entry subsystem and to the data management subsystem so that data validation can be performed. To increase the reliability of the data, it is advantageous to define the domain for each variable as narrowly as possible.

Both the data entry and the data management subsystems protect the integrity of the data by ensuring that each variable’s value is always within its domain and that duplicate records cannot be entered. Each variable’s value is defaulted to “missing.” Because several different people enter data into the BPD database and because different data-entry personnel tend to have different error rates, we constructed the database software so that each data record automatically includes the identity of the person who entered the record and the date of entry. This allows researchers interested in quality control issues to take into account different error rates when analyzing data.

Security
Two types of security are required in the database: security against accidental loss and security against unauthor-
ized access. These aspects of security are critical to the success of the database and must be provided at multiple levels.

At the first level, Ingres provides security against accidental data loss by allowing a “roll-back” of the database to its previous state when data items are incorrectly modified or deleted. The roll-back security feature is occasionally required during everyday database work to solve problems introduced by power outages and by prevention of access to a record during attempted simultaneous access. A second level of security against accidental data loss was initially provided by an automatic backup facility written into the ABF/QBF scripts, but was abandoned when the overhead of keeping additional backup copies of all relations was deemed unjustified. Ultimately, of course, security against data loss is provided at a third level by regular daily backups of the computer system to tape.

Security against unauthorized access at the first level is provided by the Unix password security features. A second level of security is offered by the system’s requirement for user authorization on a per database and per relation basis, as well as on a per field basis within each record. A third level of security, that not all users be given full access to all databases they are permitted to use, is also provided by user authorization.

Statistical Analysis

The Statit statistical package (Statware Inc., Corvallis, Oregon) provides an interface to the Ingres database management system. The package includes modules for window management, data management, graphics, descriptive statistics, and inferential statistics, such as hypothesis testing, nonparametrics, analysis of variance, regression, multivariate analysis of variance, and discriminant and factor analysis. Statit supports the use of SQL commands incorporating operations for copying data from the Ingres database into the Statit workspace. This interface can be used to perform statistical procedures on subsets of data without explicit data conversion. Our initial design incorporated the Statit package as an integral part of our database structure.

If a researcher needs to use another software package, data can easily be exported to an ASCII file. For example, when we needed to use an unbalanced repeated-measures technique for analysis of longitudinal data, we downloaded the data in ASCII and transported them to the BMDP 5V program for analysis.

User Interface, Data Entry, and Reports

The data entry protocol for the BPD study specifies an intermediate paper-based stage because the logistics of the study preclude direct, online entry of data as they are collected. To minimize errors and data entry time, forms were constructed directly from those in the software, with very little manual work required. The paper forms serve as an additional level of backup to ensure data integrity.

The user interface was built around the Application-by-Forms (ABF), Query-by-Forms (QBF), and Vision facilities of the Ingres software. The interface is user-friendly and easy to modify. The ABF facility permits many of the “automatic” features of the database (e.g., backups and calculations) to be easily incorporated and modified as required. The program runs on a variety of computer terminals without modification.

The user interface for the BPD database was built as a four-tiered hierarchy. At the top is a main menu offering business/medical information, interview data, test data for the infant subjects, and test data for their parents. At the second level are the menus for these four categories. Each of these second-level menus provides a choice of individual relations. At the third level are menus regarding the type of access: “append,” “browse,” “update,” etc. At the fourth level are customized screens permitting data entry and retrieval.

During the study, most of the data entry was performed by part-time student employees with a high turnover rate. Despite this, the students quickly learned to use the interface and overall productivity was maintained. We isolated them as much as possible from the nuances of Xenix and Unix, by making the data entry software start automatically as they logged on. Other project personnel had the usual Xenix or Unix shell access with special aliases defined to give easy access to the database.

The need for both ad hoc and regularly scheduled reports was recognized. Ad hoc reports were constructed and run in the Ingres SQL environment. Reports required on a regular basis were prepared and made accessible through the menu system. Among these are four reports used to identify registered subjects for whom data collection is incomplete, one report used to locate unregistered subjects for whom data have been collected, a report that lists projected testing times for specified patients, and a report that lists scheduled visits during a specified interval.

Status

The database was initially developed over a six-month period in 1989, and required approximately 250 person-hours of effort. The initial implementation was carried out primarily by a graduate student with no previous experience with Ingres but with a strong background in programming and database theory. Subsequent development work has been done by two different full-time employees with a strong background in programming but little or no experience with database management systems. About 700 person-hours of effort have been spent on developing and implementing the database so far.

The database has been in use since early 1990. Approximately 908 separate analyses have been performed with its data. Of these, approximately 88% were done entirely within the confines of the Ingres–Statit implementation described earlier. For the other 17% data were copied from the database and analyzed with another statistical package (SPSS, BMDP, Systat, or SAS). In these cases, logistic regression, analysis of variance, and multivariate analysis of variance with more than one covariate, which are not provided in Statit, were required as part of the analysis.
At present the database consists of 55 relations: one holds permanent demographic data, one holds variable demographic data, and the remainder hold the results of various laboratory tests. The relations that contain test data are keyed to the identification number of the patient, the sequence number of the test (1 through 5), and the familial relation to the patient (mother, father, grandmother, etc.). Users report that they are pleased with the ability to add to the database and modify relations as studies develop. They report that even the most inexperienced staff members are able to enter data successfully because the validation checks reduce entry errors, the custom menus and entry screens make the system easy to use, and the security features prevent inadvertent destruction or modification of data. Users also report that time is saved and errors are reduced by the integrated scoring procedures using calculated variables.

Conclusion

The BPD database has successfully evolved to fit the changing needs of the BPD study and has served to support analytical studies resulting in several publications [16–19]. Since creating this database we have begun developing similar research databases for other projects—on Linguistic Sequelea of Unilateral Lesions (for the National Institutes of Health), Propofol Sedation for Acutely Ill Ventilated Children (for the Food and Drug Administration), Sensory Motor Development of Cocaine-Exposed Infants (for the National Institute of Drug Abuse), and Pediatric Health Supervision to Promote Literacy (for Maternal and Child Health Services).

(From the Department of Epidemiology and Biostatistics, the Department of Pediatrics, and the Department of Computer Engineering and Science, Case Western Reserve University, Cleveland, Ohio. Address reprint requests to Toyoko S Yamashita, Ph.D., at the Department of Pediatrics, Center for Medical Informatics and Statistics, Triangle Suite 200A, 11400 Euclid Avenue, Cleveland, OH 44106 (tsy@hal.cwru.edu).

Supported in part by grants (NIH ROI-HL 38193 and MCI 390592) from the National Institutes of Health and Maternal and Child Health Services. The authors thank the members of the Infant Follow-Up Program and the Center for Medical Informatics and Statistics, Department of Pediatrics, for their support. Portions of this work were presented at a conference entitled Bedside Computing in the 90's, Park City, Utah, March 1991 [18].)

References

7. Woolf PK. Deception in scientific research: AAAS-ABA national conference of lawyers and scientists, project on scientific fraud and misconduct (report on Workshop No. 1), Washington, DC: American Association for the Advancement of Science, 1989.