Use of a Microcomputer Database System in a Statewide Effort for Data Collection in Medical Genetics

Joyce A. Mitchell, Ph.D.
James H. Cutts, M.S.
Mimi Hess, M.S.

Medical Informatics Group
University of Missouri-Columbia
Columbia, MO 65202

Abstract

The Genetics Office Automation System (GOAS) is a database management system for the collection and reporting of medical genetics data. We have previously reported on its implementation in a single university center [1,2]. We report here on its implementation in a coordinated data collection effort for the State of Missouri. We discuss the current status of the data collection activities and procedures to share data collected at an individual center with state, regional, and national data collection efforts.

Introduction

The Missouri Department of Health (DOH) and the tertiary care centers in Missouri have long recognized the need to collect data on genetics services; they began such activities in 1984. The primary foci of the data collection effort are those of program evaluation, provision of services to rural areas, and coordination of care with other state programs. Other reasons have become important in the last several years including (a) the contribution of data to the Great Plains Genetics Services Network (GPGSN), a cooperative effort within an eight state region which would assist in the assessment of services and diagnoses in a large geographic area, (b) the contribution of data to the Council on Regional Networks (CORN), a cooperative effort between regional services networks which would collate and analyze medical genetics data nationally and provide input to the national Bureau of Maternal and Child Health, (c) the desire of local centers to have better knowledge of their own clinical activities and patient population.

The data collection activities established in 1984 by the Missouri Department of Health (DOH) were in need of revision and refocusing in order to accommodate these additional data collection and analysis needs. The old system had included paper forms which were filled out by hand and mailed to the DOH. These procedures had all of the inherent problems of a paper-based system: poor handwriting, data entry by people who did not know genetics and could not easily ask the originating center for clarification, inadequate proof-reading resulting in many database errors, poor data analysis, and poor compliance from clinical centers because of lack of feedback in data analysis.

In 1990, the DOH decided to totally revamp the data collection and analysis procedures in an attempt to overcome these problems. The method chosen was to use a microcomputer database system which had been in operation since 1985 at the Division of Medical Genetics, University of Missouri- Columbia (the GOAS system [1,2]), and to install that system into all of the centers collecting data. With a concerted effort by all of the centers to concentrate on quality control and uniform collection of data, this new system has the potential to overcome the difficulties with the old system and to satisfy the additional data collection needs as well. The conversion from a paper-based system to an electronic one puts the emphasis on data entry and validation at the point of collection of the data (the individual centers) where the expertise exists to verify and proofread the data. Data extraction programs are run monthly and the data is sent on floppy disk to the DOH where it is transferred into a statewide database system for analysis. We now have quality-checked data available in a timely fashion. This data will be analyzed regularly and presented to the collecting centers, the state-wide advisory committee, and will hopefully give adequate feedback to insure the continued collection of high quality data.

0195-4210/91/$5.00

Methods and Procedures

The Genetics Office Automation System

The database management system that has been designed to collect this set of data is called the Genetics Office Automation System (GOAS). The system consists of a program written in the dBase III+ language, and the 34 databases that are used to store the collected data. The GOAS program manages data entry and reporting.

The databases in GOAS are logically divided into primary and reference databases. The primary databases contain records relating to a patient, their visits and diagnosis. The reference databases are used to validate and drive data entry as well as reduce the total system storage requirements. The records in a reference database are not directly related to any individual patient. The primary databases can be sub-divided into three sub-categories: (a) principle, (b) record keeping and (c) diagnostic.

(a) The principle databases store the core information relating to the patient and their visits. One record per patient in the *patient* database contains the patient's name, date of birth, sex and other demographic information. The patient record is the center of all the information on the patient. By following the path laid out by the pointers stored in the patient record, it is possible to access all the information on the patient. Each visit made by a patient generates one record in the *visit* database. This record will store the who, where, when, why and what of the visit. Associated with the visit record are records in the *test* database that store the tests performed or ordered in association with the visit.

(b) The record keeping databases store information that is not used in determining a patient's diagnosis, but is necessary for the identification of each specific patient. Each patient is assigned identifying numbers by various different organizations. These numbers, used to protect the patient's identity, are stored in the *numbers* database. The *contact* database is the repository for information about the people to be informed about the patient progress with provisions for each patient to have two separate contacts. The *doctor* database provides information about two doctors associated with each patient: the referring physician, and the patient's family physician. The *doctor* database can be considered midway between a primary and a reference database. It is used as a primary database containing information relating to a specific patient. It is implemented as a reference database where the information in the database is not directly related to any specific patient, and each doctor's address is only stored once.

(c) The diagnostic databases are *dx*, *findings*, and *family name*. *Dx* is the most important diagnostic database. It stores the description of the current genetic diagnosis for a patient. The categories of diagnostic information are: chromosome abnormalities, Mendelian disorders, multifactorial conditions, teratogen exposure, other recognized syndromes/ conditions, undiagnosed syndromes/ disorders and normal. Each patient can have entries in more than one of the diagnostic categories. In addition to their own diagnostic information, diagnosis for two relatives can be stored in the association with the patient. The entries in the Mendelian disorders category are selected from the reference database with the numbers and descriptions of Mendelian Inheritance in Man [3]. The *findings* database stores descriptions of key physical features present in the patient. In the next version of GOAS (under development), the findings will be selected from a controlled vocabulary modeled after the London Dysmorphology Database [4]. There are provisions throughout the system for local entries to be made to the diagnostic reference databases. The *family name* database stores the surnames of the paternal and maternal grandparent of the patient. Other information such as consanguinity and eligibility for special research studies may be kept in the family name database. The database is used to search for linkages between families previously seen in clinics.

The reference databases are used during data entry to validate the information being entered and to present a controlled vocabulary from which to choose a field entry. During data entry, if the value entered is not found in the reference database, a menu is created based on the contents of the reference database, and the data entry person can select the correct value from that menu.

By storing a pointer into a reference database in the primary databases, instead of the full text description, considerable space can be saved. The implementation of GOAS currently running in the Genetics Clinics of the Hospital at the University of Missouri - Columbia (MU) is able to store over 1800 patients with more than 6000 visits in less than 3 megabytes.

The design of the GOAS databases includes "intelligent"

reference databases. Intelligent databases are reference database which learn what is the expected input for a field that may not have specific constraints before hand. The *city-county* database is an example of an intelligent database. This database matches the names of cities with the counties in which they are located. Whenever a city is entered, the city-county database is checked. If the city is not found, the data entry person is requested to enter the county. The association of the city and county is recorded in the *city-county* database. The next time that city is entered, its record will be located and the county automatically entered. It is still possible for the data entry person to change the city and country linkage if necessary (some large cities may have residents from several counties). "Intelligent" databases have the property that the longer the system is used the less frequently the database entries are changed because all of the standard entries are included in the menus.

Most of the fields in the primary databases refer to values used repeatedly in an associated reference database. In the current version of the program, only a few reference databases (such as the *city-county* databases and the *Mendelian number-name*) have a predefined content. Currently, the contents of the intelligent databases are subjected to periodic review in an attempt to standardize the contents of the system as much as possible. In the next version of the program, most of the reference databases will have predefined content which can be chosen from a list during data entry and pasted into the appropriate field. This will help to insure that the vocabulary is controlled and misspellings are minimized. Also, the next version will include a proofreading routine for each database. These routines will automate the proofreading process by printing reports and by consolidating the updates of records in the reference databases.

Over fifty report routines are already available in the GOAS which can be used by each center. The procedures for using these reports are detailed in a user's guide, with examples of the standard report formats available for perusal. Some examples of these reports: find patients with specific diagnoses, determine the county distribution of patients or referring physicians, print a visit and test summary of a specific patient, and calculate visit statistics summaries over a given time period.

Installing the System in Other Genetics Centers

Representatives from the four centers collecting genetics data and the Missouri Department of Health gathered to discuss the data collection and needs of each center and the needs of the DOH, GPRGN and CORN. The centers agreed upon the minimal set of data to be collected and the definition of each data item. They also agreed to use the GOAS system if it were refined to collect the data upon which they agreed and to refer to their specific center.

Before the GOAS could be installed in other centers, the databases referring to a specific center had to be removed and procedures for installation in new centers needed to be developed. System documentation and a users' guide were essential. Changes determined by the state-wide data meeting were also made. Each center obtained the computer hardware recommended for the system: IBM or compatible 386 computer, VGA or EGA color monitor, 60 MB hard disk and laser printer.

Slightly different versions of the GOAS program were implemented for each center in order to fit most appropriately into the clinic operation. The GOAS was installed by a programmer who trained the users at each center in data entry and report generation procedures. The programmer travels to each center at least once a month to assist in any difficulties or questions which might arise. The centers' personnel extract the requisite data for the state and the Great Plains monthly in specially written report routines. The data is mailed to the DOH on floppy disks where it is aggregated into a statewide database. Incomplete data is referred back to each center for completion and retransmission.

Meetings of the data collection committee are held every six months to review progress and difficulties with the system. These meetings will begin to concentrate on data analysis as well as vocabulary review. The review process for the vocabulary of the databases will take place at two levels: the center level, and the state level. At the state-wide meetings the representatives will determine which local observations are being made across the state and come up with a uniformly accepted name by which to report the observation to the rest of the state.

Minimum data set

The data collected by all centers is a rather simple set of five relational files: (1) personal, (2) general, (3) visit, (4) diagnosis, and (5) tests. Although there is a great deal of overlap between the data collected for the Missouri Department of Health and the Great Plains Genetics Services Network, the data fields differ in format and method of numeric coding. The current procedure to gather the GPGSN data is to use a program already written for the GOAS which extracts data from each center and transforms it into the codes requested from the

GPGSN. This data will be transmitted by the Missouri Department of Health to the GPGSN on a yearly basis. The GPGSN, in turn, extracts the data requested by the Council on Regional Networks (CORN) for the entire eight state region and sends it to the national offices. This cooperative data collection effort makes the data from each cooperating genetics center in the United States available for use by state, regional, and national centers in a commonly defined format with identical definitions.

Results

The new methods for statewide data collection system are still in their infancy, but some results are already apparent. Overall, within four months of startup, the new data collection procedures are working smoothly. The conversion from the manual system to the computer system went relatively well, but had some anticipated glitches: hardware problems, printer and peripheral definition problems, difficulties with the correct data being available for entry, some inevitable changes needed in the software, and some desperate calls for help with data entry. The travels of the programmer to each center have been most useful in alleviating these difficulties.

As is known in most database systems, the less typing the better when entering data in order to avoid typographical errors and insure uniformity of data. We are working diligently on a revision of the GOAS so that the data entry person will select items from menus instead of typing. This is most important when one considers that there are at least 5000 different diagnostic entities which could be seen at any of the genetics clinical centers. The new system will have access to some standard databases of genetic diagnoses, including the OMIM [5] database of diagnostic numbers and names of Mendelian disorders and the same nomenclature for dysmorphic physical features as used in the London Dysmorphology Database [4]. Further, the new version will be written in Clipper which runs faster than the dBase III+ of the current system. The revised version should be complete and installed in the data collection centers by the middle of the summer.

Data summaries and statewide analysis are not yet available. The data has been collected in this standardized manner from December 1, 1990. We are working on the analysis of quarterly data summaries. We anticipate that there will be an analysis of statewide service coverage to consider underserved areas, a consideration of the most common diagnoses throughout the state, and a comparison of demographic variables with previous years. By regularly presenting these and other analyses to the centers, we hope to insure that each center feels a personal contribution to a worthwhile statewide effort.

Discussion and Conclusions

This statewide data collection system is not just a description of the current computer system and methodology, but also a description of a revolution in data collection methodology from the previous system. We know that the computer system works even in the first few months of its operation. We know that the data can be read by the state, integrated into a statewide database system, and is available months sooner than was possible with the previous system. Furthermore, data extraction routines allow for the data from the entire state to be contributed to the Great Plains Genetics Services Network in a timely fashion.

We hope that all of the potential benefits of this system can be realized. There is the possibility of creating a system where

- All data is uniform
- Data is only input once
- Data entry is the responsibility of the center which generates the data
- Feedback loops provide the originating centers with a sense of importance for their data collection activities
- Clinical data collection centers have their own data in a computerized form with report routines already written which make that data accessible to them.
- Data needed on a local, statewide, regional, and national level is collected in an easy, uniform, and accurate manner.

The model of local data being collected which conforms to state, regional, and national data definitions in a manner which is useful to all of these entities is being pioneered by the field of Medical Genetics and may be emulated by the rest of the medical profession. This model and its associated procedures could be useful for many such distributed data collection efforts beyond the field of Medical Genetics.

Acknowledgements

We acknowledge the medical centers within the State of Missouri who are collecting this data and using the GOAS program: University of Missouri - Columbia, University of Missouri - Kansas City, St. Louis University, and Washington University. This work is supported by the Missouri Department of Health contract N0A0000511.

References

1. Cutts, J.H. III, Mitchell, J.A. Microcomputer-based genetics office database system. In Ackerman, M.J. (Ed.) Proceedings 9th Annual Symposium on Computer Applications in Medical Care (SCAMC). Computer Society Press: Washington, D.C., 1985: 487-490.

2. Cutts, J.H. III, Mitchell, J.A.: The Genetics Office Automation System, a database management system for medical genetics. In Hammond, E. (ed) Proceedings of the American Association for Medical Systems and Informatics (AAMSI) Congress. IEEE Press: Washington, D.C., 1988: 175-179.

3. McKusick, V.A. Mendelian Inheritance in Man (9th Edition). The Johns Hopkins University Press: Baltimore, MD, 1990.

4. Winter, R.M., Baraister, M., and Douglas, J.M. A computerised data base for the diagnosis of rare dysmorphic syndromes. J Med Genet 1984 Apr;21(2):121-3.

5. McKusick, V. (1987). Online Mendelian Inheritance in Man [Machine-readable data file]. Baltimore: The Johns Hopkins University (Producer). Baltimore: The William H. Welch Medical Library (Distributor).