A MEDICAL GENETICS DATA BASE MANAGEMENT SYSTEM

K.W. Kang, A.D. Merritt, P.M. Conneally, J.M. Gersting and T. Rigo

Department of Medical Genetics and Mathematical Sciences. Indiana University, Indianapolis

This is publication #78-45 from the Department of Medical Genetics and was supported in part by the Indiana University Genetics Center, PHS Grant, GM 12054 and ROI MB109

Summary

A Medical Genetics Data Base Management System has been developed for genetic counseling and research in the Department of Medical Genetics, Indiana University. The system is designed to store and retrieve information collected on approximately 15,000 families seen over 14 years and helps collect information on clinical and research patients from approximately 1,500 families per year. The data base at the present time includes over 500,000 individuals and the primary information on many of these families is retrievable including family pedigree, genotyping and physical, laboratory diagnostic information.

Introduction

The rapid advances of medical science during this century have eliminated or controlled so many of the environmental determinants of diseases that the residual causes of morbidity and mortality of many individuals now depend a great deal upon one's own variation rather than the environment. It is becoming increasingly apparent that a major component of this variation is genetically determined, often by simple Mendelian inheritance. The science of medical genetics is relatively new and the number of newly discovered genetic disorders previously thought to be non-genetic has increased rapidly in the last decade. Over 2000 genetic disorders are currently listed from research findings.¹

Genetic entities vary from conditions which are relatively common such as diabetes (1 in 10) and cystic fibrosis carriers (1 in 25) to disorders which are relatively rare (as low as 1 in 1,000,000 live births). As increasingly effective diagnosis and treatments for these disorders are developed, the need for preventing genetic disease and improving the quality of life is increasingly more acute.²,³

The approach required to prevent this problem is somewhat different from the conventional health care approaches that have been employed in the past in preventive medicine. It involves, as a starting point, the detection of those individuals in the population who are at risk of having a child with a serious hereditary defect, or those who have the genetic potential of being affected by such a problem. With present knowledge, genetic counseling of high risk families and vigorous research are the practical approaches to the prevention of genetic disorders. There are certain genetic disorders that can be eliminated by specific environmental means; the problems associated with Pyloric stenosis can largely be avoided through early surgical intervention and, similarly, for phenylketonuria through the introduction of a phenylalanine-restricted diet soon after the birth of an affected child.

One of the major problems in genetic services is the attainment and storage of proper information related to individual medical and family history. We have long recognized the necessity for the development and maintenance of a medical genetic data base which would provide efficient access to accurate information. This would serve as an important part in clinical management of genetic disorders, particularly in diagnosis, estimation of risk, prognosis, follow-up and genetic counseling through record linkage.

During the past 14 years, 15,000 families have been seen in the Department of Medical Genetics for diagnostic evaluation and genetic counseling, or for research purposes. Family information, physical finding and health status have been obtained on over 500,000 individuals. Additional information such as specific diagnostic parameters, chromosome studies, various laboratory data (blood types, serum and salivary protein markers, dermatoglyphics, glucose tolerance test, lipoprotein and lipid analysis and many others) is also present on a large number of individuals.

The purpose of this paper is to describe a genetic data base management system which provides collection of pedigree information, pedigree linkage, soundex name search, pedigree drawing, and efficient storage and retrieval of medical as well as genetic information.
using the on-line computer. The acronym for the system is MEGADATS (Medical Genetics Acquisition and Data Transmission System).

Methods of Procedure

In general, an appointment is scheduled when a patient is referred to the Department of Medical Genetics for clinical evaluation or inclusion in a research project. The genetic assistant is responsible for sending a preappointment family history questionnaire, which is to be completed at home. In order to limit the size of the questionnaire and the time necessary to complete it, only information on the "nuclear family unit", consisting of the patient, the patient's spouse, children, sibs, parents, sibs and children and grandparents, is obtained. The patient is requested to return the completed questionnaire prior to his appointment in the Medical Genetics Clinic. At the time of interview, a genetics assistant corrects the information obtained by the questionnaire. Ancillary health and diagnostic information is checked with emphasis placed on a specific individual's symptoms and disease state. The following information is collected on each individual within the family: 1) surname (including maiden names of married women); 2) first and middle names; 3) birth date; 4) sex; 5) age of death if deceased; 6) marital status; 7) parents; 8) marriage partners, and other relevant information on consanguinity and birth place. Whenever indicated, blood and urine samples are obtained for chromosome and special laboratory studies. Genotyping and dermatoglyphic analysis are routinely performed.

This information is processed in the Medical Genetics data base system based on the following specifications:

(a) Family pedigree data processing - clinical data entry and access to pedigree data must be in a real time mode, and a procedure must exist to determine if pedigree data being entered duplicates or overlaps data already in the system (record linkage).

(b) Laboratory data processing - data must be stored in "raw" as well as processed form and be easily accessible, and procedures for verification and consistency checking must exist.

(c) Support of the departmental function - detailed information on individuals and families must be maintained in a manner consistent with clinical use of the data, and information must also be stored in such a way as to be useable for research and services in genetic counseling and follow-up of individuals.

(d) Security of information - privacy of all information must be guaranteed, and systems must be maintained to provide the confidentiality of data and restrict the use of all or part of the data to those who have a need for the data.

An overview of the system is presented in Figure 1. The first and most important part of the system is data collection. This block represents the source of data (the input) and, for the most part, the destination of the results (the output). As previously mentioned, the input to the system (pedigrees and lab data) is well defined and the desired output of the data processing effort is fairly well defined, so that the remainder of this section will deal only with the computerized effort to produce the output from the input.

Briefly, the structure of MEGADATS (enclosed in double lined blocks in Figure 1) consists of four modules: the control module directs the activity of the system, maintains security to prevent unauthorized access, and keeps a log of all activity for security and accounting purposes; the input module processes and stores all data collected, whether it is a pedigree or results from lab tests; the support module contains functions such as updating files, searching files and the general "housekeeping" chores that are inherent in a system of this type; finally, the S2K module maintains the data base which acts as the data archive. Note that Figure 1 is divided into two blocks; one indicates the portion residing on the DEC 11/45 computer in the Medical Genetics Computer Center, and the other portion residing on the IBM 370/158 in Data Systems and Services on the Indiana University campus at Bloomington (DS&S is the IU data processing support function).

The "two block" approach was adopted in the design of the data processing system for three reasons: speed, security and specialization. The DEC 11/45 is a medium-sized time sharing system which is designed to respond well to interactive inquiries. The IBM 370/158 is a very large system on which one can run a data base management system known as System 2000 (S2K). This data base system is designed to efficiently handle large amounts of data. The "two block" approach uses the DEC 11/45 as an index or catalogue which contains primarily individuals' names and family linkage (about 20M characters), and the IBM 370/158 S2K as an archive to house all data other than the names (about 60M characters). Quick indexing can be accomplished on the smaller machine, and the larger sorts and searches occur on the larger machine.

The linkage of family members is achieved through a set of six pointers shown in Fig. 2. An example of the linkage scheme is presented using a standard genetic symbol in which a
Fig. 1. A schematic presentation of MEGADATS

PREDECESSOR

Father  Mother

Sibling

CONTEMPORARY

INDIVIDUAL

Mate

SUCCESSOR

Offspring

Multiple Record

Fig. 2. Record linkage pointers
square indicates a male, circle indicates a female and the lines depict the relationships (Fig. 3). A sample family and linkage scheme is shown in Fig. 4. In a given record the pointers are scanned from left to right; if a pointer is unresolved (i.e. blank) the system asks the interviewer for information involving that pointer. The records are processed from the bottom (last record entered first) up, except for the first time when the proband is (affected individual) examined. The sib chains are closed loops with offspring pointers from mother and father to one sib in the chain. Each sib points to his mother and father. In the scheme used here, most special cases are handled by the multi-record (MR) pointer. This pointer is used to chain together multiple references to a single record. For instance, Sally Brown has a second legal name Sally Johnson (females are always referred to by maiden name). In this case, an AKA (Also Known As) record marked with an * is linked to record f9 using the MR pointer. Bill Williams is a second mate to Mary Smith and is linked to her through the MR pointer with an N prefix. Other items, such as adoptions and consanguinities, are processed using different variations of the MR pointer and prefixes.

It is important to note that the linked-list scheme is a precise unambiguous link to the family. It is part of the family information and is retained in the data base along with the other information. A pedigree may be continued simply by recalling it and "opening" one of the pointers, thus starting the list directed questioning process.

The next entry is LAB DATA which is processed through MEGADATS for the storage and retrieval of genotype data and many other types of data, such as data from chromosome studies and diagnosis and physical measurement of patients. As "laboratory data" for an individual passes through MEGADATS on its way to the archive, (SZK), indicators are set in that individual's Family Information File records to note that data is in the data base. The lab data processing includes provision for data verification and for consistency checking wherever it is possible and applicable.

The UPDATE of information is the key to successful operation of the entire system. It is the mechanism that moves information out of the temporary update file into its proper position in the Family Information File. It will also make appropriate copies of the information to be sent to SZK for inclusion in the archive data base. UPDATE is utilized for pedigrees and most laboratory data. The HOUSEKEEPING routines are a necessary part of a system this size. They include several kinds of "bootstrap" routines to get the system started, to convert previously existing data to the new format and several maintenance routines to keep things running smoothly.

System 2000 is a general purpose data base management system developed and marketed by MRI Systems Corporation, Austin, Texas. It provides the user with a comprehensive set of data base management capabilities. These include the ability to define new data bases, modify the definition of existing data bases, and to retrieve and update values in these data bases. Values for each record in the data base may vary in length. The user may specify without restriction which elements in the data base are to be inserted and become key fields.

Data elements in System 2000 can be accessed via a procedural programming language such as COBOL, FORTRAN or an assembly language, or the user can access the elements with the immediate access feature (CICS). The immediate access feature provides a user-oriented language with which a non-programmer may express requests for retrieval or updating the data base.

Security has been integral in the design of MEGADATS from the very start. Two approaches have been followed, protection and separation.
Fig. 4. Linkage pointers of the sample pedigree.

M . . . Mate, MR . . . Multiple Record, O . . . Offspring
S . . . Sib, F . . . Father and M0 . . . Mother

Date:

<table>
<thead>
<tr>
<th>IND. #</th>
<th>NAME</th>
<th>SEX</th>
<th>FAM #</th>
<th>REL #</th>
<th>M</th>
<th>MR</th>
<th>O</th>
<th>S</th>
<th>F</th>
<th>M0</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Doe</td>
<td>?</td>
<td>M</td>
<td>26186</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>5</td>
<td>14</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>Doe</td>
<td>?</td>
<td>M</td>
<td>26186</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Jones</td>
<td>?</td>
<td>F</td>
<td>26186</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Smith</td>
<td>?</td>
<td>F</td>
<td>26186</td>
<td>4</td>
<td>1</td>
<td>10</td>
<td>5</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Doe</td>
<td>?</td>
<td>F</td>
<td>26186</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Doe</td>
<td>?</td>
<td>F</td>
<td>26186</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Doe</td>
<td>?</td>
<td>M</td>
<td>26186</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Doe</td>
<td>?</td>
<td>F</td>
<td>26186</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Williams</td>
<td>?</td>
<td>M</td>
<td>26186</td>
<td>9</td>
<td>10</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Smith</td>
<td>?</td>
<td>F</td>
<td>26186</td>
<td>10</td>
<td>9</td>
<td>4</td>
<td>10</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Smith</td>
<td>?</td>
<td>M</td>
<td>26186</td>
<td>11</td>
<td>12</td>
<td>11</td>
<td>4</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Brown</td>
<td>?</td>
<td>F</td>
<td>26186</td>
<td>12</td>
<td>13</td>
<td>13</td>
<td>4</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Johnson</td>
<td>?</td>
<td>F</td>
<td>26186</td>
<td>13</td>
<td>12</td>
<td>12</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

In addition to required password protection on the host computers for accounting purposes, both MEGADATS and S2K have additional security involving several levels of password verification. For further security, the data have been "separated" in that all personal identifying information involving names resides in the Family Information File on the DEC 11/45 under the physical control of Medical Genetics in its computer center. The remainder of the data (test results, etc., but no names) resides on the IBM 370/158 in DS&S under the same physical and logical security as the student records kept by the registrar. A continuing task throughout the MEGADATS project is to evaluate and restructure the security procedures any time it is felt that unauthorized access, misuse or misinterpretation of any data could occur.

The extraction and reporting processes of MEGADATS are varied. Most extractions in the PDP 11/45 component are relatively straightforward, being based on either a family/individual number key or a name type key. Such extractions result in generation of reports in the form of a simple listing of the records from FIFM for the individuals in the family or in the form of a Calcomp plot of the pedigree using the same type of notation that would be used if the pedigree had been drawn by hand.

Extractions in the S2K component tend to be more complex in nature since all of the data collected (except names and some relationship data) are available in the S2K database. Results of extractions are usually destined for one of three general uses. The first is a response to a CICS request (i.e., an on line request) usually to determine information about a particular individual or family. The second is composed of a set of standard reports the largest of which is called the Genotype Listing. This application consists of a listing of all the laboratory test results (genotype data) available in the file. The output is organized based on a nuclear family, i.e., a mother, father and their offspring. Most often this output is made for all families but it can easily be requested for a single family. The third use of an extraction within the S2K component is the extraction of a certain portion of the data for use in research type analysis programs such as linkage analysis and programs that determine genetic risks.

Discussion

A genetic data base has been established for the recognition and dissemination of information about genetically determined diseases. The system emphasizes ease of use and clarity of the human-computer interface and employs the computer as a tool to aid the clinical and research work. The system provides the following features for medical care in areas of medical genetic services:

1) Individuals from families with genetic disorders will have access to the best genetic advice available from accurate information on diagnosis and prognosis, depending on knowledge of the genetic status of the individual and his family. 2) Information on heterogeneity,
identification of carrier status, pedigree history and family linkage will provide the appropriate risk of producing an affected child, which will hopefully prevent the repetition of the tragedy of an affected child in a specific family; through genetic linkage and polymorphic marker studies, carrier states for affected genes may be identified with expression of other genes.

3) The computer system will increase the recognition of familial aggregates of specific disorders and obtain more precise epidemiological data. 4) The system will provide an assortment of sporadic cases, environmental phenocopies or new mutations from familial isolated cases. 5) The system will provide invaluable data for future research on mechanisms of many genetic disorders; most importantly, it will improve health care through existing genetic research.

References


