Utility and Limitations of Genetic Disease Databases in Clinical Genetics Research: A Neurofibromatosis 1 Database Example

PATRICIA BIRCH* AND J.M. FRIEDMAN

Databases that collect clinical information on patients with particular genetic diseases can be used to investigate the clinical history of a disorder, its genetics, and genotype-phenotype correlations. A database can also serve as a valuable source of patients for studies of disease pathogenesis, variability, or treatment. We review the strengths and limitations of genetic disease databases in the context of our experience with the National Neurofibromatosis Foundation International Database (NNFFID). Genetic disease databases have been developed by individual investigators, scientific consortia, patient support organizations, and commercial enterprises. Databases vary from simple lists of affected individuals to comprehensive collections of detailed clinical and genetic information. Data may be obtained from people who volunteer to be included, systematic assessments of patients seen at participating medical centers, or population-based registries. Access to information may be highly restricted or widely available. These variables all affect the possible uses and usefulness of the data for research. Technical aspects of data entry, organization, storage, and retrieval, as well as issues related to data quality, confidentiality, and security, help determine how well a system actually functions. We discuss examples of research that have been accomplished with genetic disease databases and make recommendations regarding the organization and operation of these resources.

KEY WORDS: database; variability; genotype-phenotype correlation; NF1

INTRODUCTION

Structured collections of data are commonly called databases, registries, or knowledge bases. These terms are not used consistently in the literature, and for the purposes of this review we shall use the term genetic disease database to refer to a structured collection of phenotypic data, with or without molecular genetic data. We distinguish disease databases from locus-specific databases, which are primarily repositories of mutational information, specific to one genetic locus or a few similar loci. A recent survey of 94 locus-specific databases found that less than half had any associated phenotypic data [Claustres et al., 2002]. When phenotypic data are collected in locus-specific databases, only a few clinical features are usually recorded. Therefore, the difference between genetic disease and locus-specific databases is that the former focus on the disease, i.e., the phenotype, while the latter focus on the genotype. Because of this difference, genetic disease databases usually include cases that result from locus heterogeneity or mosaicism, while such cases may be excluded from locus-specific databases. Genetic disease databases may also be developed for conditions that are not caused by mutations at a single locus—such as Down syndrome or multigenerational families with a particular multifactorial condition.

Attempts to identify disease-causing mutations created a new need for collecting and organizing groups of patients with Mendelian diseases, and the advent of personal computers and user-friendly software has made the development of genetic disease databases much easier. More recently, the Internet has enabled databases to collect information from widely distributed contributors and provide information to more users, creating both new opportunities and new problems. Genetic disease databases may be developed by a single laboratory, a consortium of interested researchers, a patient support organization, or a commercial enterprise.

We shall consider various database models in this paper and review our experience with clinical databases for neurofibromatosis 1 (NF1). We shall use...
these examples to discuss database design, sources of data, accessibility to the research community, and ways in which these factors influence use of a genetic disease database for clinical research.

**GENERAL CHARACTERISTICS OF GENETIC DISEASE DATABASES**

The key challenge in setting up a genetic disease database is to make it easy to use by both contributors and researchers. The design and content of a database and its ease of use create constraints upon its accessibility, the range of research that can be performed, and the database’s flexibility for application to problems that were not considered at the time of its initial development. We suggest some guidelines for database structure and content while recognizing that, for practical reasons, some compromises are inevitable.

In general, data quality tends to be inversely related to the amount of data that is collected for each case, the detail required, and the number of different contributors to the database. Strict adherence to common policies and protocols can improve data quality, but contributors’ motivation is the most important factor in assuring both data quality and continued participation. Issues like funding, academic credit, usefulness in patient care, and the ability of contributors to initiate and participate in research activities are critical to the success of a genetic disease database.

**Database Structure**

Genetic disease databases usually employ one of two basic designs—flat file or relational. Each patient is typically represented in a flat file by one record. This can be visualized as a single row in a table. A one-page spreadsheet is a common example of a flat file. In contrast, relational databases are modular. They can be thought of as many separate tables, each with one or more rows (records) per patient. All of the rows in all of the tables that pertain to any particular patient are linked together by a unique identifier. Table I and Figure 1 illustrate the differences between these two designs.

It is important to understand that the data entry format and report formats for a database are not necessarily indicative of whether the underlying structure is flat or relational. Either type of database may have a simple linear data entry mechanism or a complex branching system that includes many different forms or screens. Similarly, some standard reports from the two kinds of databases may be identical. Nevertheless, the differences in how information is stored in flat file and relational databases have important implications for how the data can be used and the kinds of studies that can be done.

A key aspect of database design is how longitudinal data are stored. Although flat files can store a new record (row of data) each time a patient is evaluated over many years, it is more common for such databases just to revise a patient’s record with new information as it becomes available. If this is done, the original data are lost, precluding longitudinal analysis of disease progression over time.

A modular relational design is more flexible and efficient for most genetic disease databases. Data that do not change over time, such as patient identifiers and most demographic variables, are entered only once and linked to data that may be different at each clinic visit, vary in frequency of collection, or be applicable to a different extent in different patients. For example, a patient seen over several years may have one set of demographic data and a single mutation analysis but several records of growth measurements, complications, and imaging test results. In addition, there may be a need to collect detailed information on each pregnancy, although males and children have no pregnancy-related data and some women have data on many pregnancies. Relational databases are much better suited than flat files for storage of data of these kinds.

Each record in a table of a relational database must be linked to all other records that refer to the same person in all other tables. This linkage may be done in various ways (one-to-one, one-to-many, many-to-one, conditionally,

**TABLE I. Part of a Flat File Database**

<table>
<thead>
<tr>
<th>ID</th>
<th>Last name</th>
<th>First name</th>
<th>Gender</th>
<th>Date of birth</th>
<th>Diagnosis</th>
<th>Date last seen</th>
<th>Café-au-lait spots</th>
<th>Axillary freckles</th>
<th>Inguinal freckles</th>
<th>Lisch nodules</th>
</tr>
</thead>
<tbody>
<tr>
<td>03-001</td>
<td>Smith</td>
<td>Joan</td>
<td>F</td>
<td>MAR 28, 1956</td>
<td>NF1</td>
<td>JAN 15, 2003</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td>03-002</td>
<td>Jones</td>
<td>Sam</td>
<td>M</td>
<td>DEC 6, 1997</td>
<td>NF1</td>
<td>JAN 15, 2003</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>03-003</td>
<td>Lam</td>
<td>Joe</td>
<td>M</td>
<td>AUG 8, 1976</td>
<td>NF1</td>
<td>JAN 17, 2003</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>92-256</td>
<td>Joe</td>
<td>Beth</td>
<td>F</td>
<td>OCT 11, 1991</td>
<td>NF1</td>
<td>FEB 10, 2003</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>N</td>
</tr>
</tbody>
</table>

*Each patient is represented by one record (line in the table). The record may contain many more data items (columns) than shown. (All data in this table are bogus.)
etc.) and can be exploited to minimize duplicate data entry. It is also possible to include certain variable responses automatically as default values so that data entry is eliminated if the default value is confirmed. This is frequently done, for example, to attach dates to important clinical findings so that the emergence and progression of problems can be studied.

A relational database design is advantageous in several other ways as well. The modular design allows easier modification and expansion of the database. For instance, an additional module can be added for laboratory or imaging results that were not initially included, or a molecular genetic module can be added after the disease locus has been defined.

Flexibility in data entry is an additional advantage of the relational structure, as illustrated by an example from the National Neurofibromatosis Foundation International Database (NNFFID): The database consists of 7 demographic questions that must be completed at the first clinic visit and 28 core phenotype questions that are completed on all individuals at each clinic visit. The core questions cover the most important features of NF1. About 75 other data items are available but are optional. They are arranged in modules for specialty areas such as dermatology, orthopedics, ophthalmology, neurology, or developmental psychology and allow more detailed information to be collected if desired. This approach encourages the collection of a core set of high-quality data on a large number of individuals while allowing centers to customize their particular data collection needs and interests. Additional modules can be added to accommodate special studies or therapeutic trials. For example, an additional NNFFID module was developed to describe plexiform neurofibromas in great detail for a clinical history study of these tumors.

This flexibility has a cost, of course. Relational databases are generally more complicated to design than flat files. It may not be worth the additional effort...
required to implement a relational design if the database is very simple and likely always to remain so. For this reason, many locus-specific mutation databases are implemented as flat files.

Data Sampling

The information stored in a database may be cross-sectional or longitudinal. The data may be population based, but more often they are obtained from a sample of patients that is selected or self-selected in some way. The nature of the data collected should reflect the purpose of the database.

Most genetic disease databases are cross-sectional and derived from patients seen in specialty clinics because this approach is relatively easy to implement. The NNFFID is primarily this kind of database, although about one-third of the data are longitudinal. Because specialized clinics tend to attract many patients with a particular disease, clinic-based databases can become quite large. The NNFFID currently includes data on more than 4,500 people with NF1, a group that is large enough to permit study of uncommon features and to test hypotheses regarding associations between various disease manifestations.

The NNFFID currently includes data on more than 4,500 people with NF1, a group that is large enough to permit study of uncommon features and to test hypotheses regarding associations between various disease manifestations.

For example, the large size of the NNFFID enabled us to demonstrate that some NF1 patients are more likely than others to have certain serious disease manifestations [Friedman and Birch, 1997b; Baser et al., 1999; Szudek et al., 2000, 2003; Szudek and Friedman, 2002].

Data obtained from specialty clinics and regional referral centers are usually biased because more severely affected patients tend to be referred to (or seek out) such centers. This is a particular problem in disorders like NF1 that have a highly variable phenotype. On the other hand, data from specialty clinics and regional referral centers may underestimate the prevalence of disease manifestations that have high risk of mortality soon after diagnosis: affected patients die while patients with milder disease may be seen repeatedly in the clinic. For example, malignant peripheral nerve sheath tumors are often underascertained in NF1 patients in cross-sectional data because most affected individuals die at a young age and within a few years of diagnosis of the malignancy [Evans et al., 2002].

A database may attempt to capture all cases of a disorder within a geographical region. This population-based approach makes it more likely that the data collected will not be biased toward patients who seek medical attention because they have more severe disease. Birth defects registries are familiar examples of population-based data. Population-based data are especially useful for estimates of prevalence, mutation rates, reproductive fitness, and age-related incidences of disease manifestations. Longitudinal population-based data can also be used to calculate life expectancy and lifetime risks for developing particular features. Data obtained from specialty clinics and referral centers can be compared to population-based data to determine the degree and type of biases inherent in the former. Table II illustrates some of these issues by comparing the frequencies of several manifestations in NF1 in specialty clinic-based databases and population-based studies. Data sets such as those of Samuelsson and Axelsson [1981] and Poyhonen et al. [2000] include patients of all ages in a distribution that reflects the patient population as a whole. In contrast, the NNFFID includes a disproportionately large number of children because most of the specialty clinics that contribute data are located at pediatric hospitals. It is therefore not surprising that the average age and the frequency of features like discrete neurofibromas, which are more common in adults than in children,

<table>
<thead>
<tr>
<th>TABLE II. Comparative Frequencies of Five Manifestations of NF1 in Population-Based Studies and Data From Specialty Clinics</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Patient ascertainment study</strong></td>
</tr>
<tr>
<td><strong>Number of patients</strong></td>
</tr>
<tr>
<td><strong>Mean age (years)</strong></td>
</tr>
<tr>
<td><strong>Six or more Café-au-lait spots</strong></td>
</tr>
<tr>
<td><strong>Discrete neurofibromas</strong></td>
</tr>
<tr>
<td><strong>Symptomatic optic glioma</strong></td>
</tr>
<tr>
<td><strong>Pseudarthrosis</strong></td>
</tr>
<tr>
<td><strong>Seizures</strong></td>
</tr>
</tbody>
</table>

*Some patients did not have information available for all manifestations.
are lower in the NNFFID. However, it is more difficult to obtain population-based data than hospital- or clinic-based data, and this is reflected in the smaller sample sizes of most population-based studies. The estimates obtained for smaller samples are statistically less reliable, and this may produce nonrepresentative frequencies for some features in small studies. The high frequency of seizures in the Samuelsson study may be an example of this effect.

Although ascertainment bias is inherent in data obtained from specialty centers, some kinds of studies are unlikely to be seriously affected by this bias. Examples include analyses of associations between clinical features such as axillary freckling or Lisch nodules in affected individuals or family members [Szudek et al., 2000, 2002, 2003]. These clinical features are unlikely to influence ascertainment.

In addition, the availability of data on a large number of patients in a genetic disease database may provide a valuable means of identifying particular subgroups for more detailed study. For example, the NNFFID was used by Lin et al. [2000] to find NF1 patients with cardiovascular malformations, by Stevenson et al. [1999] to locate patients with tibial dysplasia, and by Gutmann et al. [2002] to identify adolescents and adults with brain tumors for further investigation. Use of the NNFFID greatly facilitates the identification of patients for studies of uncommon manifestations of NF1.

Use of the NNFFID greatly facilitates the identification of patients for studies of uncommon manifestations of NF1.

A further consideration in data sampling is the need for a stringent definition of the disease for case inclusion. The inclusion criteria for the NNFFID are the clinical definition of NF1 developed by the National Institutes of Health (NIH) consensus conference [1988]. Cases may be coded as NF1 or possible NF1 based on these criteria, but only definitive cases are included in cross-sectional analyses. A consistent definition of the disease is necessary for replication and comparison of analyses. However, at some stage in clinical research for most diseases, the clinical definition is not clear. A database can be used to help define these clinical criteria. In neurofibromatosis research, an illustrative example is the pioneering study of Crowe et al. [1956]. This was the first major systematic examination of a large series of patients with multiple neurofibromatosis. This study suggested criteria for the clinical definition of NF1 and enabled the authors to estimate the disease prevalence and mutation rate.

Familial Data
Each of the three severe manifestations of NF1 included in Table II has a higher frequency in probands than in affected relatives in the NNFFID. This illustrates the importance of identifying the proband in a database in which ascertainment is not population based. Patients with more severe (but not immediately lethal) disease are more likely than those with mild disease to be seen in a specialty clinic or referral center. In databases collected from such centers, the frequency of serious manifestations in affected relatives (i.e., nonproband) probably provides a better estimate of the true population frequency than the frequency in probands.

Collection of familial data and precise identification of the relationships of affected relatives to each other are essential for studies of intra- and interfamilial variability. Such data can be used to determine which features are likely to be influenced by allelic heterogeneity, by modifying genes, by variation in the normal allele, or by chance [Szudek et al., 2002].

Coding familial relationships is theoretically simple. A standard method consists of assigning each parent of every affected individual in the family a unique identifying number and including these numbers in the record of the individual. Sibs can then be defined as subjects who have the same maternal and paternal numbers; half sibs are subjects who share one parental number but not the other; cousins are subjects who share grand-parental numbers; and so forth. A separate question is needed to identify the proband within each family. This numbering system permits a standard pedigree to be constructed from the data stored in family members’ records (Fig. 2).

Data Quality, Quantity, and Detail
Data quality is of prime importance in any database but is a particular concern when clinical information is collected from many different contributors. Differences among contributors in knowledge of the condition, training, health care practice, language, and culture may affect the consistency of data collection, interpretation of clinical findings, and data coding. In a genetic disease database, it is important to confirm the diagnosis of every case using standard clinical criteria or reliable testing. The NNFFID has a check box to identify type of NF, but we occasionally find patients designated as NF1 who do not meet the standard diagnostic criteria for clinical diagnosis [Gutmann et al., 1997]. However, the clinical information collected in the NNFFID permits independent confirmation to ensure that patients selected for studies of NF1 actually fulfill diagnostic criteria.

Data quality is enhanced by use of a data dictionary that defines every item and permissible response in the database. For example, Lisch nodules are defined as present in the NNFFID only if they are seen on slit lamp examination.

Logical response and range checks can also be used to identify data entry errors. Logical response checks use combinations of responses that are inconsistent or very unusual to help avoid data entry errors. For example, coding a subject as a male who has an age at menarche of 13 is a logical inconsistency that could generate a data entry error message. In the NNFFID, a pop-up warning message appears if a child is
coded as having an age of 17 years and as being prepubertal or not yet in school. Range checks determine whether the value entered for a quantitative variable lies within an expected range. For example, a data entry warning will appear in the NNFFID if the height is coded as less than 50 cm or more than 200 cm.

A common source of error in data contributed by centers from many different countries is inconsistency in units of measurement. Data entry errors are minimized by consistent use of metric units for all measurements, but many centers in the United States do not employ metric units. In the NNFFID, an option exists for data entry in imperial units. Data that are entered using this option are converted and stored in metric units by the software.

Date formats are a particular concern in international databases. 03/02/01 usually means March 2, 2001, in the United States, but in other countries it means February 1, 2003, or February 3, 2001. We require all dates to be entered into the NNFFID in standard international format (four-digit year, two-digit month, two-digit day), but we automatically convert them to an unambiguous written format that is displayed adjacently: thus a date entered as 2001-04-12 is also displayed as April 12, 2001.

Language differences may be an issue in databases. We wrote the NNFFID program in English only, and it is a credit to our collaborators’ linguistic abilities that we have received excellent data from Japanese, Russian, French, Italian, Spanish, Finnish, Swedish, Norwegian, and German centers. Our database allows free text comments that are sometimes written in English but often in other languages. This can be a problem when looking for particular symptoms, features, or treatments, but searching by common Greek or Latin roots often provides an effective solution.

Clinical features are frequently recorded as binary (present or absent) variables in genetic disease databases. Although such coding has the virtue of simplicity, it limits statistical analysis of the data. Recording variables in a quantitative fashion (e.g., number of café-au-lait spots or age at onset of hearing loss) usually permits use of more powerful statistical methods for data analysis. The greater statistical power of quantitative data has proven to be especially important in the analysis of familial correlations of phenotypic features [Easton et al., 1993; Zhao et al., 2002; Palmer et al., in press].

Few clinicians take the time to count exact numbers of features such as cutaneous neurofibromas or café-au-lait spots. Use of variables with a categorical or semiquantitative response (e.g., none, 0–9, 10–99, 100–999, or \(\geq 1000\)) may provide a practical compromise in such circumstances.

An important determinant of data quality is data quantity. In general, it is harder to maintain high data quality if a larger number of items must be coded for each patient or if more detail is required for each item. The NNFFID was originally implemented with 100 obligatory data items, but we later rearranged the items into 28 obligatory core data items and a number of optional questions. Our subsequent experience shows that the core items in the revised database are sufficient for most research questions that could have been answered with the more extensive data set, and the consistency of data coding is better for the smaller set. The core data items are not sufficient for detailed studies of particular aspects of NF1, but the
database can be used to locate patients for these kinds of studies. The detailed information required can then be obtained through review of local medical records, surgical reports, imaging studies, and laboratory results. This two-step approach provides better quality data on specific issues and encourages appropriate collaboration and sharing of authorship among investigators and clinical contributors [Stevenson et al., 1999; Lin et al., 2000; Gutmann et al., 2002].

Another factor affecting quality and quantity of data is funding. We have been fortunate to receive funding for a part-time curator for the central database, but there has been no funding for contributing centers. It is likely that more complete data might be obtained if contributors also receive funding. One area of weakness in the database that might be improved by funding contributors is that of long-term follow-up: for example, we are sometimes unaware that a patient has died because entering such information is not a priority for a busy clinician.

Molecular Genetic Data

As mentioned previously, most locus-specific mutation databases contain no data on phenotypic variability [Claustres et al., 2002], and those that do include clinical information usually collect very little. Genetic disease databases often include information on mutations to facilitate detailed genotype-phenotype correlation studies. The international standards that have been established for coding and classification of mutation data in locus-specific databases [Antonarakis, 1998; Scrivener et al., 1999, 2000] should be used for such data in genetic disease databases as well.

Confidentiality, Control of Data Access, and Ethical Review

The Internet provides an opportunity to make databases easily available to contributors and users throughout the world, but very few genetic disease databases have taken advantage of this opportunity. This probably reflects general concerns about the confidentiality, security, and potential for abuse of data that are available over the Internet.

The NNFFID is accessible over the Internet, but we assure confidentiality by including no names, addresses, or local record numbers in the database. Each patient is assigned a unique NNFFID number, and only the center that contributed the data on that patient can link the record to his or her identifying information. This approach may cause problems when a clinician retires or relocates because the link may be broken. We have decided to live with this problem in order to protect patient confidentiality.

Policies on data access and use for research were established for the NNFFID at the time of its inception. Our policy permits any legitimate NF researcher to use the database. However, access is password protected by center-specific user accounts, and a user from one center cannot alter data contributed by another center. An additional confidentiality measure prevents viewers from displaying the exact date of birth on any patient not contributed by their own center. All users may perform certain types of searches, e.g., selecting patients with certain phenotypic features, but more complex searches and analyses require involvement of the database administrator. Access to detailed information or specimens on any individual patient must always go through the clinician who contributed the case. Policies have also been established regarding publications that make use of NNFFID data, and written agreements regarding ownership of the data and software were established before the database was implemented.

Each contributing center must meet the requirements of its local Institutional Review Board (IRB). Most IRBs require assurances regarding confidentiality, informed consent and assent, voluntary participation, and permission to recontact, but other issues often arise as well. In the United States, recent implementation of Health Insurance Portability and Accountability Act (HIPAA) regulations may further limit data collection for genetic disease databases.

CONCLUSIONS AND RECOMMENDATIONS

There is no perfect generic design for genetic disease databases because there are so many different genetic diseases and their phenotypic features vary so greatly. Nevertheless, we believe that the following recommendations provide useful guidance for the development of such databases:

- Policies regarding confidentiality, security, participation, data access, data ownership, and authorship related to the database should be established at the beginning and agreed upon by all participants.

Policies regarding confidentiality, security, participation, data access, data ownership, and authorship related to the database should be established at the beginning and agreed upon by all participants.

- The purpose of the database should be clearly established, and the design should fit the purpose.
- A modular relational database design allows for great flexibility of data entry.
and retrieval and for easier maintenance and expansion to accommodate future needs.

- Data quality is enhanced by brevity, range and logic checks, and a data dictionary.
- If data are collected over time on individual patients, the database should store the information in a longitudinal fashion, rather than replacing older data with newer data on a single record.
- A web-based interface provides inexpensive worldwide access but brings additional concerns regarding data use, security, and confidentiality.
- Avoiding collection of names and other patient identifiers is a useful means of assuring that confidentiality is not compromised even if the database’s security software and protocols are breached.
- For genetic studies and for studies of disease severity, progression, and complications, it is necessary to understand ascertainment biases inherent in the data.
- The proband in each family should be identified, and all family relationships should be recorded.
- Quantitative data should be collected when feasible. Semiquantitative or categorical data are more informative than binary (present or absent) data.
- It is usually preferable to design a briefer database that captures only what is essential and permits patients or families with specific characteristics to be identified for separate detailed studies.
- The key challenge in developing a genetic disease database is to make it both easy to use and useful. A balance must be achieved between the amount and detail of the information recorded and the ability of contributors to provide it reliably, consistently, and on a continuing basis.

When in doubt, choose the simpler option.

For more information about the NNFID, please contact Patricia Birch: birch@interchange.ubc.ca.

REFERENCES


