STATISTICS IN MEDICINE
Statist. Med. 2000; **19**:1531–1544

Balanced centralized and distributed database design in a clinical research environment

Norman L. Foster[1,*,†], Eszter Gombosi[2], Cheryl Teboe[2] and Roderick J. A. Little[2]

[1] *Michigan Alzheimer's Disease Research Center and the Department of Neurology, University of Michigan, Ann Arbor, MI 48109, U.S.A.*
[2] *Michigan Alzheimer's Disease Research Center and the Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, U.S.A.*

SUMMARY

Clinical research databases can meet both research and clinical needs, but this ideal is seldom achieved. Priorities often differ for those who collect and ultimately use the data and those who develop data systems. Traditional database designs also create logistical barriers that hamper communication. The Michigan Alzheimer's Disease Research Center has developed a secure, distributed data system with centralized data entry that provides an intuitive, individually customized interface for investigators in their clinics, laboratories and offices. Data are kept in a form that can be readily understood without reference to a codebook. Investigators can modify and query their own copies of the database without knowledge of programming languages. Balancing centralized and distributed designs for research databases enhance the accuracy and completeness of data collection and increases the use of data for research and clinical care. Copyright © 2000 John Wiley & Sons, Ltd.

1. INTRODUCTION

There is often a tension between those who collect and use research data and those who are entrusted to assure its security, confidentiality and integrity. No tension exists in small studies where an investigator maintains full control of the data and determines all procedures. However, in large multi-disciplinary or multi-centre investigations, data management requires a significant effort and becomes a separate endeavour. In these circumstances conflicts can develop. In multi-centre clinical trials where cases are contributed for a fee, conflicts are usually limited to the complaints about the number of data forms and the difficulty in using them. Investigators have little or no say as to what data are collected and are obliged to conform to the requests of those who have paid for the data collection. In collaborative research that uses data contributed by independent investigators pursuing their own research, data collection is potentially more contentious. Investigators must provide to others data they view as their own. Furthermore, each discipline and centre may have

*Correspondence to: Norman L. Foster, Department of Neurology, University of Michigan, 1920 Taubman Center, 1500 E. Medical Center Dr., Ann Arbor, MI 48109-0316, U.S.A.
†E-mail: nlfoster@umich.edu

Contract/grant sponsor: NIH; contract/grant number: P50-AG08671

different views about appropriate use of data. With the loss of full control of their own data, many feel threatened by the possibility that data might be misused or misappropriated by others for a purpose they have not approved. They may find that others have taken advantage of their work without providing proper credit. These feelings are reinforced if investigators are unable to readily retrieve and review the collected data themselves, or if the collected data are difficult to correct, use in clinical care or obtain for analysis.

Data managers, on the other hand, have a different perspective. They must be concerned with standardization of procedures, efficient data management and data security. They may become insulated from the concerns of those collecting and wishing to use the data. They may find investigators reluctant to share data and resistant to recommended procedures. Although they ultimately share the same goals, their different concerns may draw investigators and data managers into an adversarial, rather than co-operative, relationship. These conflicts are increased further if data collection procedures or forms do not reflect research procedures or interfere with patient interactions. Traditionally, data managers had few options, and databases were fully centralized to address the concerns of data managers despite investigator complaints. Data were coded to conform to the requirements of computer software and hardware. This centralized and inflexible structure of research data systems largely has persisted, even though technology now exists for a distributed design with improved accessibility for users.

Alzheimer's Disease Centres (ADCs) provide an excellent opportunity to utilize innovative database design to promote collaborative research. ADCs have been developed with funding from the National Institute on Aging (NIA) to improve the understanding and treatment of Alzheimer's disease and related disorders. Centres fund research and support individual projects through shared, centralized core resources utilizing patient information, including an Administrative Core, Clinical Core and Neuropathology Core. Some ADCs also provide other shared research cores, depending upon their individual priorities and interests. Each ADC has an Executive Committee that determines its priorities and use of resources.

ADC structure is based upon the recognition that collaborative, multi-disciplinary research is required to address the complexity and broad biological, human and social ramifications of dementing disorders. Data obtained by the cores are shared among all the cores of the centre and with individual projects. Core leaders and core staff are often involved in research projects themselves, and thus have a potential conflict of interest. Cores and research projects are often multi-disciplinary and require data sharing among investigators that may be simultaneously competitors and collaborators. For example, characterization of patients by the Clinical Core utilizes data from laboratory studies, brain imaging, neurologists, psychiatrists and neuropsychologists. Final diagnosis also requires the expertise of a neuropathologist. Clinical correlates of basic science observations require that data from an investigator are collated with data from the Clinical and Neuropathology Cores. Enhancing communication, facilitating data collection and data sharing between these diverse groups is a demanding task requiring careful planning, extensive consultation and tact.

In addition to data sharing within an individual ADC, all ADCs are obliged to share data with the Alzheimer's Disease Data Coordinating Center for access to a wider audience of researchers. This collaboration adds further complications, because there is no consensus among dementia researchers about which clinical measures are most useful. Unlike most disorders, such as cancer, heart and renal disease, there are no established blood tests or other biological markers available for the diagnosis or to rate the severity of Alzheimer's disease [1,2]. Consequently, all ADCs differ somewhat in the methods they use for evaluation and are encouraged to explore different research approaches and develop new clinical measures. Each centre provides unique strengths and interests to the field, but this diversity is also a challenge to data managers. Some centres primarily involve neurologists, others primarily psychiatrists. Some use a physician-centred model of clinical care, others use a nursing-centred model. These differences are reflected in different research approaches. Centres have developed their own data collection procedures in response to their individual interests. Innovative approaches will be needed to establish a shared collection of reliable data from all ADCs, while not stifling new clinical inquiry. Rigid prescription of all clinical data collection that can be reasonably contemplated in other fields is not warranted at this stage of Alzheimer's research development. Therefore sufficient conformity is needed to assure comparable and useful information from each centre, without having shared data procedures being so burdensome that innovation is discouraged.

The Michigan Alzheimer's Disease Research Centre (MADRC), is one of 14 NIA-funded ADCs that funds both individual and pilot research projects. It is located at the University of Michigan in Ann Arbor and also maintains two satellite diagnostic and treatment centres (SDTCs) elsewhere in Michigan, one at Harper Hospital in Detroit to serve urban and African American patients, and the other in Northern Michigan at the Munson Medical Center in Traverse City to serve rural patients. Each SDTC uses the same measures as the University of Michigan site, but also has developed some unique procedures to meet the needs of their particular staff and patient population. For example, the Detroit SDTC is monitoring family responses to requests for provisional consent for autopsy to see whether different procedures will encourage participation by African Americans and the Northern Michigan team includes a pharmacist who closely monitors use of prescription and non-prescription drugs. Consequently, the MADRC encompasses the combination of conformity and diversity in approach that is characteristic of ADCs. The MADRC has developed a separate Biostatistics Core to provide investigators support for data management and analysis. Biostatistics Core staff are responsible for overseeing data collection in Ann Arbor and work closely with designated staff at each SDTC to assure appropriate collection and transfer of data from these remote sites. It provides a single repository for shared data and plays a pivotal role in reviewing the activities of the centre. To meet the concerns of both clinical researchers and data managers, the MADRC has developed incrementally over the past 10 years a multi-user fully relational database system that integrates the advantages of centralized and distributed designs. This report will present the strategies for the design and implementation of this database and its effects on clinical research and patient care.

2. DATABASE DESIGN

Data management for all research supported by the MADRC is the responsibility of the Biostatistics Core, and is a resource available to all investigators. The core assists researchers in the design, conduct, analysis and interpretation of investigative studies. Data management is used to help achieve these objectives, with the recognition that sharing information enhances research and is cost efficient by avoiding duplication of effort and conflicting information. A single relational database has been developed that integrates patient and normal control subject information from all sources – clinical interactions, research studies and post-mortem examinations. Data are also included about caregivers and family members. Patients in the database are cross-referenced to a published citation when their data has been reported in the scientific literature. This confluence of data maximizes potential efficiencies, but requires significant co-ordination and communication between database personnel and investigators.

Not all information available about a patient or developed by a research project is entered into the research database. A limited amount of clinical information is recorded. Items are included if they are: (i) relevant to the completeness and results of the diagnostic evaluation; (ii) needed for screening subjects for research studies, (iii) useful for contacting patients, their families and their health providers; or (iv) of research value to other investigators. The database is not meant to duplicate or substitute for the medical record and computerized information provided by the health system for all patients, such as laboratory results, appointment records and billing.

Patient data developed in the course of research studies are an important component of the database, but not all research data are included either. Researchers maintain ultimate responsibility for the data they collect, but are expected to share information with the database if they receive direct funding from the MADRC or in-kind support through its core facilities. This expectation is explicitly stated in agreements between the investigator and the MADRC Executive Committee, and in letters of support. When appropriate, it is expected that database activities will be described in project applications for funding, in submissions to the institutional review board, and in consent forms. Investigators are encouraged to provide as much data as possible to the database, but the items that are eventually included in the shared data set are determined by a joint decision between the MADRC core leaders, the data manager and the investigator. In some cases, it is advantageous for investigators to utilize the shared database for all data management, however this is usually not possible. Longitudinal data, data collected by more than one research project, and data from multidisciplinary studies are particularly suited for inclusion in the MADRC database. An agreement can be made that some of the data incorporated into the MADRC database will be available only to a specified and limited group of individuals. This provides investigators with a significant degree of control over their own data and encourages data sharing by helping to address investigator concerns about the misuse and misappropriation of their data. Investigators also can obtain assistance in the design of a database for their own project. While they are completely responsible for the entry and management of data in this unshared individual database, this arrangement permits the database to be designed so that items can be transferred easily to and from the shared database.

We realized that investigators would be collecting a large number, sometimes hundreds, of variables on relatively few individuals. Furthermore, for most subjects, the same data elements would be collected repeatedly in longitudinal studies and during visits for clinical care and could come from many locations. We therefore needed to have a fully relational database. Since error checking in a centralized database is best when it is close to the point of data collection [3], we wanted the data to be accessible to those providing it. We also recognized that data entry would be more complete and accurate when the database itself enhanced the daily clinical and research needs of those providing the data. Consequently, we needed to provide an easily accessible and up-to-date data collection in which investigators could benefit from the shared data and be able to readily evaluate its accuracy.

3. DATABASE IMPLEMENTATION

3.1. Location and distribution of database

To meet its objectives, the database is stored in the offices of the Biostatistics Core, but is also available to users on personal computers. Helix RADE software for Macintosh operating systems (Single Helix Corporation, San Diego, CA) is placed on personal computers in clinics where patients receive their clinical evaluations, and in the laboratories and offices of MADRC investigators and staff. This software was chosen because it provides a customizable interface and intuitive, graphical data queries that maximizes accessibility of the database to users with varying degrees of familiarity with computers. Data can be easily uploaded from external files or downloaded from the database via universal ASCII files into spreadsheets, statistical and graphics programs for additional and more complex data analyses.

The master copy of the database file resides in the Biostatistics Core and back-up copies are archived weekly and stored in separate fireproof facilities. The Biostatistics Core distributes copies of the database to the network of personal computers. Individual users may use copies of the database on computers at several locations, if it is more convenient for them. Copies of the database are distributed weekly to personal computers through a campus Ethernet network. Updates of the database for the Detroit and Northern Michigan SDTCs are dispatched on high capacity disks by overnight delivery. The SDTCs have access to the Internet, but we continue to use disks because thus far we have found current transmission rates too slow and telephone connections by modem too undependable at our satellite sites for automatic and consistent electronic transfer of the large (approximately 50 Mb) database files. Users may archive their current copy of the database, otherwise individual copies of the database on the distributed network of personal computers are overwritten with each update. Thus, data integrity is achieved while providing current information to each investigator and for each patient contact.

3.2. Database content and security

The structure of the database has been continuously modified by the data manager (E. Gombosi) in response to the needs of its users and to maximize its efficiency. The database itself currently consists of 22 sets of interrelated data tables or relations, each of which is linked to the others by common patient identifier fields and represented graphically in the program by an icon. The database also defines what information can be accessed by individual users. Each user is indicated graphically in the database program by a separate icon. The specifications for each user icon provide password protection for database access and determine the format and extent of data elements available. Users can both read and modify (write on) their personal copies of the database, but such modifications are only temporary since they are not on the master copy and are overwritten weekly. The ability to modify contents of the database is helpful to users because it can aid them in reporting corrections and allows customizing non-recurring searches and reports.

To maintain confidentiality, access is only given to data forms and reports that the user can justify for purposes of patient care or research to the satisfaction of the data manager and those providing the data for sharing. The well-defined user access privileges to specific data and password security are reassuring to investigators who can determine which individuals are able to read their data. For example, only data on psychological test performance that are useful for patient screening are available to research study nurses, while more extensive psychological test results can be provided to investigators for their research after approval of the neuropsychologists providing this information to the database. Likewise, results of genetic testing that have significant legal, ethical and financial ramifications are available only to investigators after approval of the Clinical Core Director to avoid references to these confidential results in the clinical record or inappropriate release to families.

Along with individual user passwords, an additional password is needed to view or modify database structure. After entering this level of the database structure, simply pointing and clicking on the program icons reveals specifications of the database. Thus, investigators with the help of the data manager can easily check how the database handles access to their data.

Because copies of the database are updated weekly by overwriting existing versions, only changes of data structure on the master copy of the database are permanent. The Biostatistics Core alone has access to the master copy of the database through password and access protections on their computer. To maintain data confidentially, individual users must keep their passwords secret, close the database program when not in use, and limit access of others to their computers with the database. Additional programs can be installed that require a password for access to the computers after a fixed period of inactivity, providing further security. With the co-operation of users, the use of these multi-level passwords assure a reasonable level of patient confidentiality and data security, even though copies of the database are distributed to many personal computers.

There are seven kinds of components to each data table or relation, indicated graphically in the program by a distinctive icon (Figure 1).These types of components are: (i) data elements or fields; (ii) data elements calculated from other data elements; (iii) data entry and report forms; (iv) blueprints for these forms; (v) indexing rules; (vi) sorting rules, and (vii) instructions for posting information from one relation to another. Individual data elements are the essential elements of the relation and the only items that are directly entered. The other types of components make these data fields more useful. For example, forms make the data elements accessible in a variety of formats, indexing enhances performance and sorting rules provide more useful data presentation. The 28 components of the smallest relation in the MADRC database containing data from magnetic resonance imaging studies are shown as icons from the program screen display in Figure 1.

The characteristics of individual data elements are defined in consultation with those who provide and use the data. Data is entered and displayed in ways that are most convenient for clinicians and investigators. Since the program allows text of any length to be stored, entries need not be coded and can be easily reviewed without reference to a codebook. The value of entering and displaying text rather than coded data cannot be overestimated. Although not often a major consideration for data managers, use of clearly understood text makes information readily accessible to non-technical users and enhances goodwill and collaboration. If needed, data can be coded during export to other programs or for specific views in the database through calculated data elements. Text of unlimited length and content is particularly useful when storing clinical reports and interpretations. Definition of data elements also can be restrictive with built in checks and validations. Specific types of responses can be required such as yes or no, a date, a number within certain limits, or a choice from a menu of several pre-set responses.

3.3. Use of the database

Data are recorded on forms designed by users to reflect the way they collect data. Forms used for entering or listing the same data may appear quite different, depending upon personal preferences of users. This decreases the number of forms needed to record data, and data entry personnel are responsible for translating the data to the format required by the structure of the database.

The Patient Information Sheet, shown in Figure 2, is an example of a form that is used both to record and report patient data. It illustrates how both clinical and research needs can be met simultaneously. This form collates and summarizes information on a single patient that has been entered in several different database relations from a variety of sources. The Biostatistics Core provides clinicians with a copy of this Patient Information Sheet before each expected patient

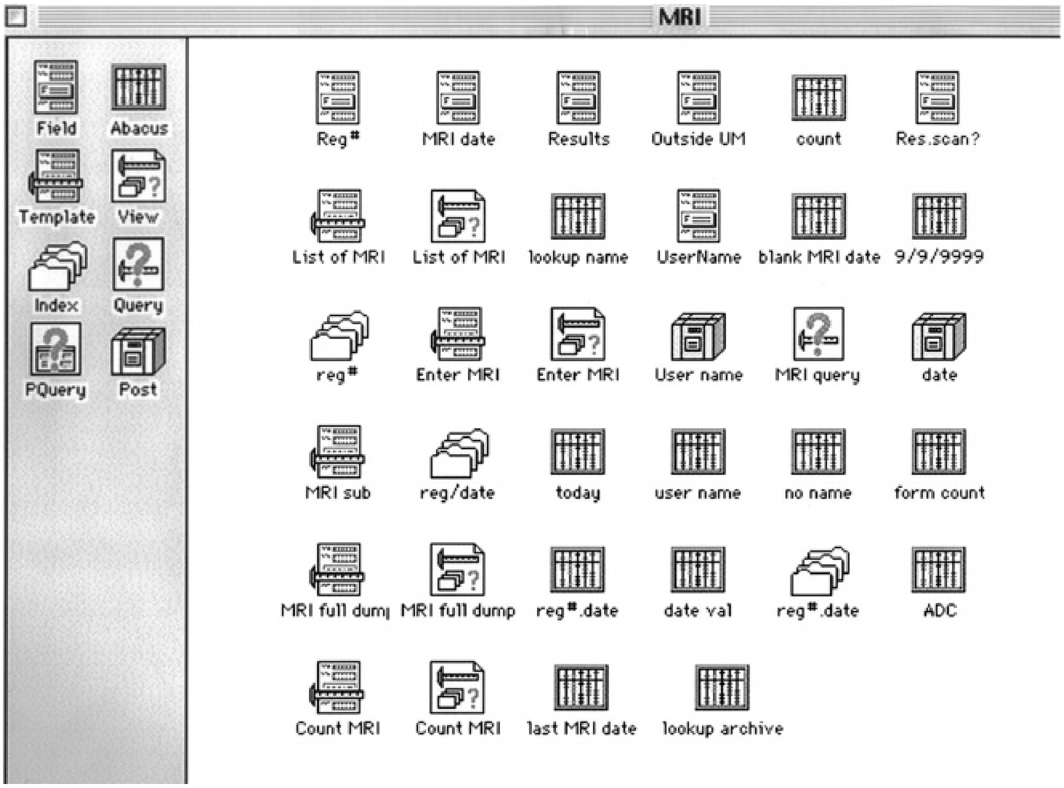
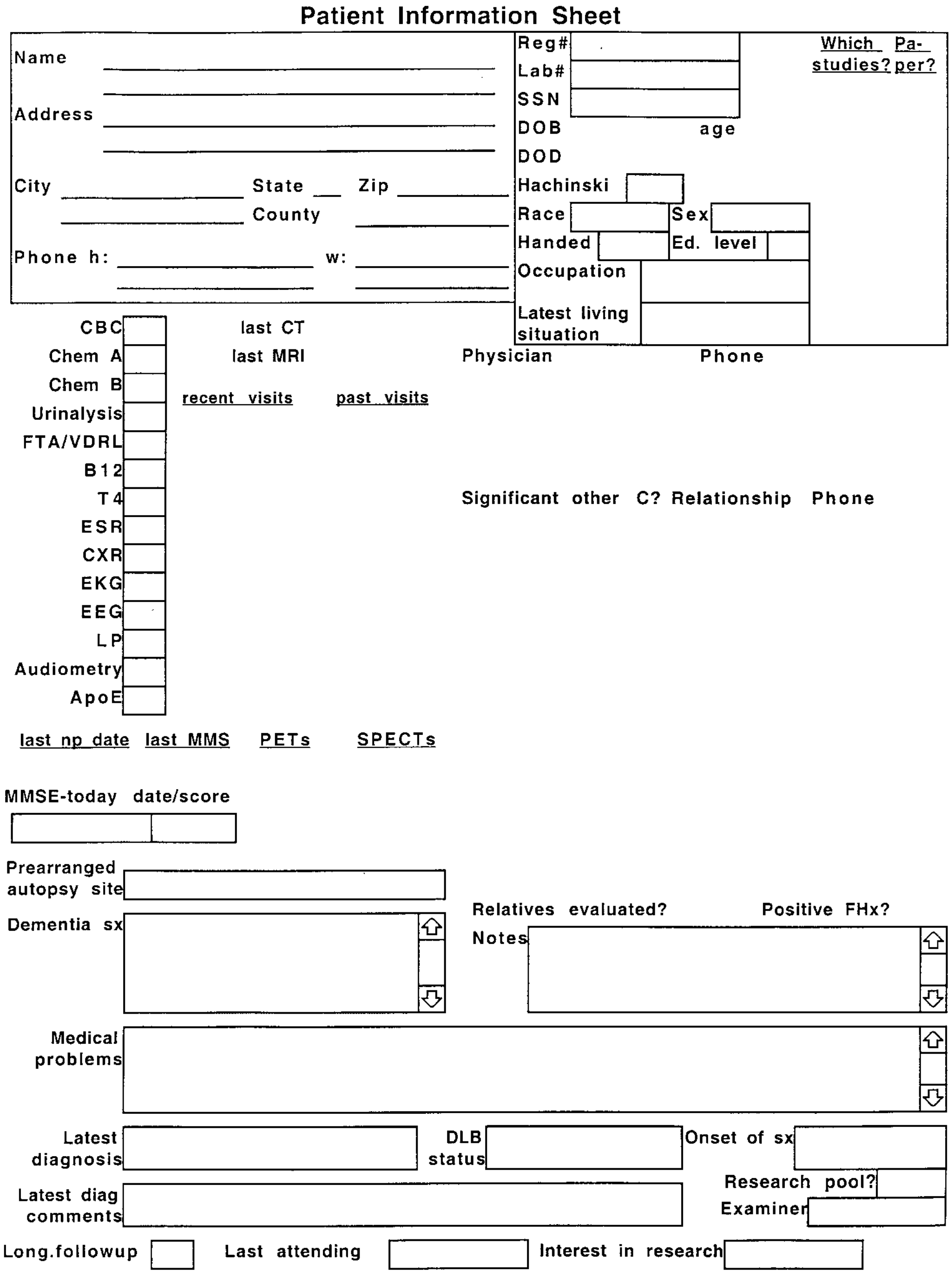


Figure 1. A computer screen from the MADRC patient database displaying the components of the MRI relation or data table. The left side of the picture indicates the icons used to build a relation, each specifying components of a particular type. The Field icon identifies individual data elements. The Abacus icon indicates data elements calculated from other data elements. The Template icon is used to create blueprints for forms. View icons generate snapshots of data entry and report forms. Index icons indicate rules for indexing or sorting data elements. Post icons are instructions for moving data around in the database, while with the Query and Pquery icons simple or more powerful queries of data can be generated. Each icon can be opened by the click of a mouse to reveal more detailed specifications.

contact. It is also often printed out directly from personal computers in the network and viewed on the computer screen during telephone calls. When referring to this summary form, clinicians are reminded of items that are helpful in patient care, but can also be utilized by researchers. Notice that the column on the left of Figure 2 provides a list of tests commonly included in a dementia evaluation [4]. These data elements indicate whether the test was done, rather than specific test results. Intended to provide information about the completeness of the evaluation, abnormalities are noted in the 'medical problems' section further down on the form. Assessing the completeness of evaluation serves as a quality check for the clinician and is useful to investigators screening for clinical studies. In the middle of the right side of the form, names of the primary care physician, caregivers and family members are listed for quick reference, also indicating relationships and those who serve as guardians and surrogate decision-makers. This information is helpful in both clinical and research contacts with patients and their caregivers.



Patient Information Sheet

Name ______________________
____
Address ______________________
____
City ____________ State __ Zip ________
____________ County ________
Phone h: ____________ w: ____________

Reg# □
Lab# □
SSN □
DOB age
DOD
Hachinski □
Race □ Sex □
Handed □ Ed. level □
Occupation □
Latest living situation □

Which studies? Paper?

CBC □
Chem A □
Chem B □
Urinalysis □
FTA/VDRL □
B12 □
T4 □
ESR □
CXR □
EKG □
EEG □
LP □
Audiometry □
ApoE □

last CT
last MRI

recent visits past visits

Physician Phone

Significant other C? Relationship Phone

last np date last MMS PETs SPECTs

MMSE-today date/score
□ □

Prearranged autopsy site □

Dementia sx □

Relatives evaluated? Positive FHx?

Notes □

Medical problems □

Latest diagnosis □ DLB status □ Onset of sx □

Latest diag comments □ Research pool? □
Examiner □

Long.followup □ Last attending □ Interest in research □

Figure 2. A Patient Information Sheet, one of the forms that summarizes data in the database. This form displays information about a single individual. It is one of the forms that can be viewed by users from the copy of the database residing on their personal computer. Brief entries such as dates or abbreviations that appear on this form reveal when additional information is available on other linked forms and reports in the database. Triangles indicate entries that are selected from a pull down menu of pre-set responses.

Some parts of the form indicate what specific types of clinical and research data are available for that patient so that other forms can be accessed for details. For example, dates of positron emission tomography (PET) and single photon emission computed tomography (SPECT) scans are listed. The clinician can then focus his record review or refer to database forms describing the results of these tests, hence avoiding fruitless or incomplete searches. During the patient's visit clinicians add and update information when appropriate on the Patient Information Sheet, which is then returned to the Biostatistics Core. There the data are entered using forms designed for each database relation. All of these data report features save clinicians considerable time and enhance the quality of patient care. With this system incomplete and erroneous data are as objectionable to the clinician as they are to the researcher or data manager. Thus, all of these three components of the research team become partners in data maintenance.

The Patient Information Sheet also directly aids research. A column in the upper right hand corner of the sheet is used to list research studies in which the patient participated. Study nurses can quickly identify demographic data that can exclude or include subjects. Clinicians report patient interest in research as part of a separate clinic contact form and this appears in the lower right hand corner. Thus research issues and status are apparent to the clinician at each contact and recruitment of patients into studies is enhanced. Limited information about the severity of dementia (MMS and MMSE on the form standing for Mini Mental State Examination score) and the timeliness of the available data help investigators perform an initial screen for possible research subjects. Another data list, the Monitor Study Form (not shown), summarizes a patient's recruitment, screening and participation in all research studies and accompanies the Patient Information Sheet for clinic visits. It collates information about whether a patient qualifies, is interested in participating, has agreed to participate, is actively enrolled or has completed involvement in a research study. The summary also indicates the reasons for disqualification, whether there was early discontinuation of the study, and group allocation, if the study has been unblinded. This allows the clinician to meaningfully discuss research with the patient and to suggest other studies, knowing that they are not in conflict.

Notice also that the form collates information used by both the Neuropathology and Clinical Cores. Through this form, updated information obtained in one core is provided and easily available to the other. The status of autopsy pre-arrangements on the left side of the form and information on caregivers and the number and relationship of family members found on the right side of the form are essential to the successful completion of post-mortem examinations. Records of deceased patients are maintained in the database and this form quickly identifies the date of death (DOD) for such individuals and other forms developed for Neuropathology Core data can be examined for details.

3.4. Centralized data entry

A distributed data system provides the opportunity for direct data entry by anyone. This is attractive because it reduces the physical and intellectual separation between data collection and data entry. Our experience, however, suggests it is best to use dedicated personnel for centralized data entry. This decreases the demands on clinicians, helps assure the standardization and completeness of data, and clearly establishes lines of responsibility. We now exclusively use central data entry, accomplished by a single person who devotes her full time effort to this task (S. Teboe). Data entry forms list the person providing the data and the database program automatically documents

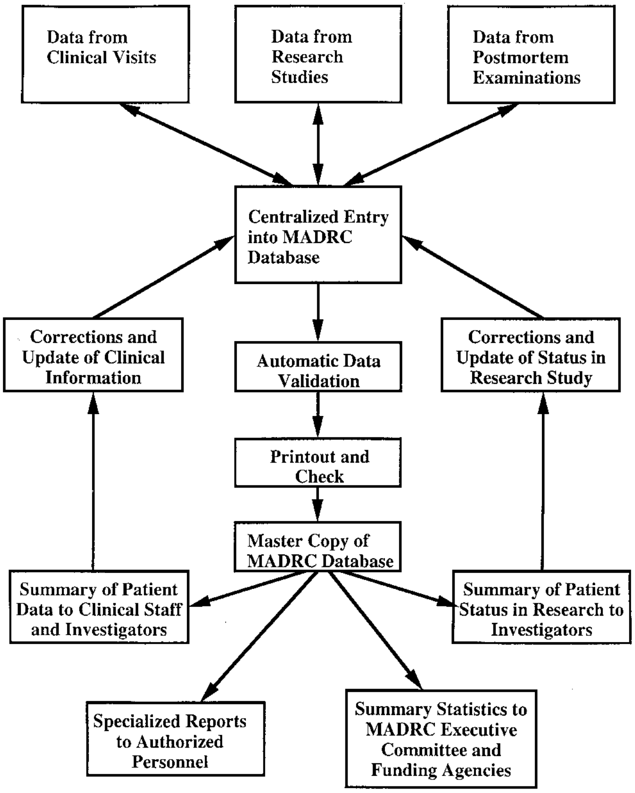

Data from Clinical Visits
Data from Research Studies
Data from Postmortem Examinations
Centralized Entry into MADRC Database
Corrections and Update of Clinical Information
Automatic Data Validation
Corrections and Update of Status in Research Study
Printout and Check
Master Copy of MADRC Database
Summary of Patient Data to Clinical Staff and Investigators
Summary of Patient Status in Research to Investigators
Specialized Reports to Authorized Personnel
Summary Statistics to MADRC Executive Committee and Funding Agencies

Figure 3. A diagram of the flow of data between the database personnel, clinicians and research investigators.

the date and identity of the person entering the data into the computer. Despite the centralization of data entry, the flow of information is bi-directional and very interactive (Figure 3). There is timely addition of new data and incorrect or out-of-date data are quickly identified and changed. Clinicians sometimes complete blank forms and other times modify already completed forms containing data that are currently in the database. Data can be submitted by writing directly on the forms (currently the most popular method) or by entering data and printing the form with the modified content. Biostatistics Core personnel collect forms daily and enter the information into the master copy of the database. Data are usually entered within 48 hours. Data personnel return submissions they find ambiguous or incomplete and check to make sure that expected forms have been completed and returned.

3.5. Data audits

Data audits are the responsibility of the Biostatistics Core staff, core leaders, and project investigators. Summary reports on individuals, such as the Patient Information Sheet (Figure 2), readily reveal missing data. Systematic searches for missing and out-of-range data elements are performed by the Biostatistics Core and addressed by investigators. Searches often are focused on subsets of patients identified by their participation in specific studies as recorded in the database.

Centralized data entry provides several strategies to assure complete data submission. Reference to the clinic appointment system alerts data managers when data forms should be received for clinical visits. Research investigators are likewise responsible for timely data submission. Data managers use periodic reports and summaries generated from the database and reports required by the MADRC Executive Committee to alert investigators that additional data is expected according to their protocol. These close interactions between investigators, data managers and data entry personnel are essential for establishing a collaborative relationship and to assure data integrity and completeness as each evaluates and supports the work of the other.

3.6. Documentation and training

The relationship between users and data managers is well defined and depends upon extensive documentation of data collection and management procedures that now extends more than 90 typewritten pages and is available on all personal computers with the database. The original documentation and each subsequent change are the result of a consensus of users and data managers. Developing the precise language in the documentation is essential to reaching consensus and understanding between data managers and database users. The final documentation communicates the consensus to everyone and is handy for everyday reference. Misunderstandings that arise between users and data managers while each is attempting to follow these agreements are investigated and are the impetus for most changes in data procedures and the documentation. The detailed description of rules and procedures help data entry personnel interpret and correctly enter data that appear on returned Patient Information Sheets, Monitor Study Forms and other notes.

Although the data management program is powerful and complex, it is possible with a limited amount of instruction for individual investigators and research staff to use the database in their everyday activities, and even to develop their own queries and reports. Data managers take the responsibility for providing training sufficient to satisfy the needs and interests of the users. Most investigators and staff who have frequent patient contacts find the database indispensable and are receptive to training opportunities.

Although individual investigators do not have the power to directly modify the master database, they are free to modify or use their own copy of the database. Lists and reports in the database incorporate a method to perform simple queries. New forms and calculations can be performed with an intuitive, icon-driven procedure that does not require the use of programming language and can be taught in about an hour. An example of a calculated data element is shown in Figure 4. Since the individual database is overwritten with each update, these changes are temporary, unless the personal copy of the database is archived. However, this archive will not reflect updated data that is entered into the master database. If the investigator wants to save these changes in updated versions of the database, he or she must send a printed copy of the changes to the data manager.

Although the training described here requires significant efforts by data managers, these training efforts are more than repaid. Educated users reduce the data manager's work by following procedures they now better understand and by decreasing the demand for simple data reports. Requests for lists and reports also are better thought-out and developed. The data manager then can focus on tasks that she is best prepared to perform such as uploading subsets of data from other databases, downloading or recoding data for analysis, designing forms and displays of data, and preparing especially complex or non-recurring special reports.

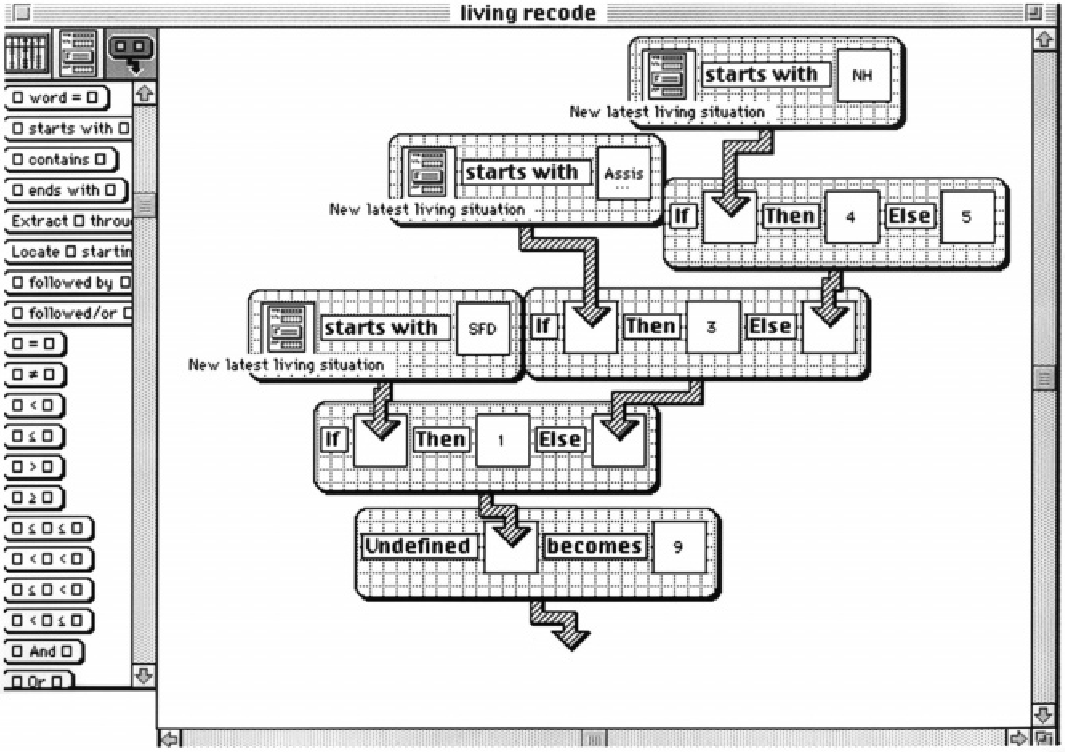


Figure 4. A computer screen from the MADRC patient database displaying the specifications of a calculated field (abacus) and demonstrating how the database uses intuitive icons dragged from an extensive list of options on the left side of the screen instead of programming language. In this example the data element 'Latest living situation' is queried and recoded for transfer to the interim Alzheimer's Disease Data Coordinating Center that uses coded rather than textual entries. All entries starting with NH (used to indicate nursing home in our database) are recoded as a 4. Assisted living is recoded as 3, all types of SFD (single family dwelling) are recoded as 1, other entries are recoded as 5 and if the field is empty it is coded as 9. The user can easily check results of the calculation in a separately designed report form. This also illustrates the advantages for users of using easily recognized text rather than arbitrarily coded entries.

4. DISCUSSION

Balancing centralized and distributed database designs has worked well for our centre, providing flexibility for users with accurate, timely, consistent, secure and confidential data. The attention and high priority given to data management by the MADRC leadership is critical to the success of the data system. Data have been provided from essentially all of the more than 60 research studies that have been supported by the centre, involving more than 1700 of the approximately 7000 subjects in the database. Although a few of these studies have added additional data fields, most have utilized and contributed to already established fields. MADRC leadership also has provided support for continuous development and improvement of the database. Too often, once a database has been developed it is expected to function without further investment. This is inadequate for a complex undertaking such as an ADC, and a high priority given to data management is essential for successful collaborative multi-disciplinary clinical research carried out at several sites.

Although it is important in the clinic, a clinical research database is not a clinical information system in the usual sense. It can and should be used to supplement, rather than replace, other sources of patient records, such as hospital or outpatient databases. There are important similarities, but also differences in the requirements for a clinical research database and a clinical information system, and the designs of these databases must also be different. Hospital data systems generally provide a unidirectional transfer of information such as retrieval of laboratory results or transmission of doctor orders. The ideal research database should promote a bi-directional dataflow. Data entry in clinical information systems may be distributed, but usually occurs in only one location for a particular type of information. We believe central data entry is better for a research database. The demands for data in clinical information systems are predictable and relatively static. A research database should never be static and must adapt to new research projects and to research advances. In both cases, significant investment is needed, even if it is not apparent to everyone that data issues are fundamental to the enterprise. That only a few successful models of a complete and integrated clinical information system exist despite great expenditures and thousands of hospitals and clinics utilizing computers attests to the challenges of the task of providing for patient care. A research database therefore should not undertake to also incorporate all the information that is needed for clinical care. Nevertheless, an effective clinical research database can still enhance the information that is available to most clinical investigators when seeing patients and prove to be a significant asset for them and thus earn their full support.

Considerable effort is required in most centralized databases to follow up on data errors and to correct outdated information. This often occurs in the form of data audits performed exclusively by data personnel [5]. Such audits are usually so effort intensive and expensive that they can be performed for only a minority of the registered cases. They also have the disadvantage of usually being limited to a review of the written record and do not directly involve those who originally collected the data. This is understandable because the audits are usually not done until long after the original observations were made, and, even if the investigators and staff are available, audits are often perceived as confrontational. By comparison the highly interactive design of our database has clear advantages. At the MADRC audits are performed by both the Biostatistics Core staff and by investigators and clinicians in the course of their work. Clinicians can review data in the context of clinical interactions, and make timely corrections that benefits data users and data personnel alike. Continuous communication between all personnel is important and should be part of a continuous joint data management effort, and not occur only during specified database audits.

Training investigators and staff who collect data about the data system is also critical [6]. As investigators develop experience with the database, they find it increasingly useful. When data collection becomes a mutually beneficial activity, it also becomes self-sustaining. MADRC data personnel have gained a reputation for being helpful and sensitive to the needs of researchers. They are able to provide advice to investigators about the management of their own data and simultaneously help assure the expansion and improvement of the shared data set. Training users also permits data personnel to concentrate their efforts on complex and demanding tasks and centralizing data entry, while making it possible for investigators to perform simple data analyses on their own. As a result, investigators feel empowered and take ‘ownership’ of the data. Without the burden of menial tasks, data managers are able to attend to details and make enhancements in the database that otherwise would be deferred or never accomplished.

By recognizing the needs of clinical researchers and their staffs, centralized data personnel can become collaborators instead of adversaries and the quality of data can be improved. Data systems balancing centralized and distributed database designs offer significant advantages for clinical research databases such as those needed for multi-disciplinary, longitudinal research on dementing disorders.

ACKNOWLEDGEMENTS

Support for this work was provided to the Michigan Alzheimer's Disease Research Center by the National Institutes of Aging (NIH grant P50-AG08671).

REFERENCES

1. The Ronald and Nancy Reagan Research Institute of the Alzheimer's Association and the National Institute on Aging Working Group. Consensus report of the Working Group on: 'Molecular and biochemical markers of Alzheimer's disease'. *Neurobiology of Aging* 1998; **19**:109–116.
2. Foster NL. The development of biological markers for the diagnosis of Alzheimer's disease. *Neurobiology of Aging* 1998; **19**:127–129.
3. Austin DF. Cancer registry types: goals and operations. *Aging* (*Milano*) 1990; **2**:80–83.
4. American Academy of Neurology. Practice parameter for diagnosis and evaluation of dementia. (summary statement): report of the Quality Standards Subcommittee of the American Academy of Neurology. *Neurology* 1994; **44**:2203–2206.
5. Levy PS. Data base management for a registry of dementing diseases. *Aging* (*Milano*) 1990; **2**:222–226.
6. Scherr PA. Cancer registries: methods and procedures. *Aging* (*Milano*) 1990; **2**:83–88.