© The Japan Society of Brain Tumor Pathology 1997

ORIGINAL ARTICLE

Sandra A. Mamrak · Allan J. Yates

An information system to support collaborative brain-tumor research

Received: October 30, 1997 / Accepted November 7, 1997

Abstract We have developed and are currently maintaining a Neuro-Oncology Information System (NOIS) to support brain-tumor research. The system has been adopted for use by a group of researchers at several centers in the U.S.A. who are collaborating to input, store, and retrieve information concerning tissue specimens, corresponding clinical and diagnostic information, and eventual research results. The information system consists of a database and user-interface components. The database has clinical, diagnostic and tissue-inventory subsystems. the user interfaces allow for input and query functions on the database.

Key words Brain-tumor information system · Tissue-inventory data · WWW-based access · Medical database

Introduction

Each year in the United States there are approximately 18000 new cases of tumors affecting the central nervous system, and the incidence of high-grade gliomas is apparently increasing. Survival for patients with some of the less aggressive types, such as cerebellar pilocytic astrocytomas, is quite good, but the clinical outcome for patients with more aggressive tumors is very poor, especially for older patients. Although it is clear that there is a critical need for new diagnostic and therapeutic modalities for these tumors, there are many difficult problems facing investigators in this area. First, the biology of these tumors is extremely complex because of the cellular and molecular heterogeneity of even specific diagnostic types. Second, because of the location of the tumors, there are frequently limited amounts of material available for studies, and serial sampling of specimens is rare. Third, survival is strongly correlated with the age of the patient and the location of the tumors necessitating control of these variables for any outcome studies. Fourth, because of the incidence of these tumors, very few centers treat large numbers of patients with newly diagnosed brain tumors. For these reasons, there is a serious need for investigators to share well-characterized specimens of brain tumors from patients with detailed follow-up. However, to do this effectively requires a powerful information system that can store and retrieve several types of information easily and can function over the Internet. We have specifically developed such an information system to overcome the most serious information problems in interdisciplinary research on human brain tumors.

After briefly describing the background of the project, we discuss the information system from four perspectives: the data needs of researchers, the design specifications for the system stemming from the unique data needs, the actual implementation of the information system, and the requirements for users of the system.

Background

The Neuro-Oncology Information System (NOIS) is being maintained to support various research projects that are pursuing the goal of improved diagnosis and treatment of patients with brain tumors.[1,2] To date, it is providing diagnostic, clinical, and tissue-inventory data to two specific groups of investigators: 1) the Brain Tumor Program Project at The Ohio State University and 2) the Glioma Marker Network. The latter involves groups of investigators at six universities in different cities in the United States. For both groups, research is facilitated by the cooperative exchange among their members of ideas and resources such as brain-tumor tissue and blood, clinical information, and research results.

The main source of data in NOIS is from patients who suffer from brain tumors and who participate in the research. There are three major kinds of treatment for brain tumors: radiotherapy, chemotherapy, and surgery. The latter usually involves the removal of neoplastic tumor tissue,

S.A. Mamrak (✉)
2015 Neil Avenue Mall, Department of Computer and Information Science, The Ohio State University, Columbus, OH 43210, USA
Tel. +1-614-292-2770; Fax: +1-614-292-2911
e-mail: mamrak©cis.ohio-state.edu

A.J. Yates
Division of Neuropathology, The Ohio State University

and this tissue is used by investigators for research. Radiotherapy and chemotherapy do not generate tissue, but because they affect clinical progression, information related to these is collected. In addition to data about these three clinical treatments, an inventory of available tissue is kept, as well as histological diagnoses. The researchers subscribe to a method of consensus diagnosis to improve the quality of the diagnosis. Several pathologists give their own diagnosis and then agree on the consensus diagnosis.

Data needs of collaborative researchers

The design and implementation of NOIS were driven by the data needs of collaborative brain-tumor researchers. The data needs can be best understood by considering research activity in which first a hypothesis is put forth, then an experimental design is planned, and finally data analysis is done to test the validity of the hypothesis.

Data for experimental designs

Brain-tumor tissue must be shared among institutions to provide sufficient numbers of tissue specimens with a given diagnosis for testing research hypotheses. Without these numbers, researchers would not be able to achieve statistical significance in their tests. A typical question that a researcher may ask at this stage of the work is, "For cases with a diagnosis of glioblastoma multiforme, are there any for which we or a collaborating member has frozen tissue or RNA available?"

This general class of queries requires that researchers have access to data on inventories of tissues, RNA, and DNA. In addition to learning of availability, they will also want to know the exact location of the tissue in the freezer or in the molecular pathology laboratory where the DNA or RNA extraction was performed. The neuropathology data about the tissue, i.e., its histology subtype and grade, must be able to be linked with the molecular pathology data so that the diagnosis is known for RNA or DNA samples.

There also needs to be a means of reconciling possibly multiple tissue identifiers, because typically the neuropathology and molecular pathology departments will have their own, distinct tissue-naming conventions. For example, the neuropathology department may precede their name with the abbreviation NP and then may add numbers that indicate the order of the tissue in this year's set, or the date on which the tissue was procured. Similarly, the molecular pathology department may use the preface MP and then assign a global, increasing number to each specimen.

Data for analysis and for testing hypotheses

When the researchers move into the analysis phase of their research, they need clinical and diagnostic data to correlate with their research results so that they can test their hypotheses. A typical question at this stage is, "For those patients whose tissue I used in my experiments, what are the relevant clinical data, including data on survival?" The relevant clinical data typically include the birth date, surgery data, and status of preoperation chemotherapy or radiotherapy. Survival data also are of interest, indicating the current status of the patient and how long the patient has maintained his or her status since surgery.

This general class of queries requires that clinical data be maintained and constantly updated until the patient's death, for each patient in the study, by the collaborating institution that provided the tissue. It also requires that the quality of the stored data be as high as possible, i.e., that it be consistent and complete. For example, if follow-up on a patient lapses by more than 6 months, data analysis might be adversely affected.

The consistency of calculated (as opposed to merely stored) data must also be as high as possible. The semantics of the calculated data must be agreed to and understood by each researcher, and the calculation itself must be done correctly. For example, a calculation of survival time is performed by subtracting the surgery data from the date the person died or was last known to be alive. This calculation can be ambiguous if the patient had more than one surgery. It is also prone to incorrect values if the input person substitutes the day the last follow-up data were obtained for the day the patient died or was last known to be alive.

Data for design and analysis

In both the experimental-design phase and the analysis phase of a research project, the correct diagnosis of the tumor is of the utmost importance.

Diagnosis of brain tumors is done by pathologists who view slides, and it is subject to differences in individual interpretation. A diagnostic error adversely affects a researcher who is trying to design a statistically sound study. For example, when a researcher queries various tissue inventories for numbers of tumors with a particular diagnosis, the researcher must be confident that the tumors in the inventory have the correct, desired diagnosis. Similarly, diagnostic errors adversely affect data analysis. For example, if a researcher is testing if a marker can be used to distinguish two types of brain tumors, and the tumors have not been correctly diagnosed, the analysis will be meaningless. For high-quality brain-tumor research, a diagnostic procedure that supports a consensus diagnosis among the best pathologists in the world is essential. Data on the results of this diagnostic activity must be available to researchers.

Design specifications for an information system

The various data requirements described in the previous section dictate the design of an information system to support brain-tumor researchers.

Database

A centralized database to store the data that researchers require is a necessity. The database must be designed to contain tissue-inventory, clinical, diagnostic, and possibly research-data subsystems. The database should support a flexible ordering for inputting the data into various subsystems, because, especially in a collaborative project, tissue-inventory data may be available well before clinical data are provided. Or the opposite may be true. The database should be designed and maintained to achieve the highest level of completeness and correctness of the data. Care must be taken to preserve patient confidentiality.

Easy access to the database

Collaborative brain-tumor researchers are scattered among institutions throughout the United States. The data in the information system must support remote data input and queries for these researchers and their staffs. Multiplatform use of the system must be supported, because researchers will be using Macs, PCs, or even Unix workstations to access the information system. Finally, access to the system must be such that it accommodates the firewalls that are found at most medical institutions.

Ease of migration of datasets

During the data-analysis phase of their studies, researchers may want to download data from the central store and combine it with their own private research datasets for analysis. The information system must support an export/save function to the researcher's local disk. Even more important, the system should support an aliasing mechanism for tissue so that system and researcher identifiers for tissue can be easily matched. For example, a researcher may want to assign his own name for the tissue in the central store so that he can do a query on clinical and diagnostic data only for tissue having his alias. Then he might want to download these data, identified in the same way in his private dataset, and merge the two datasets for the final analysis.

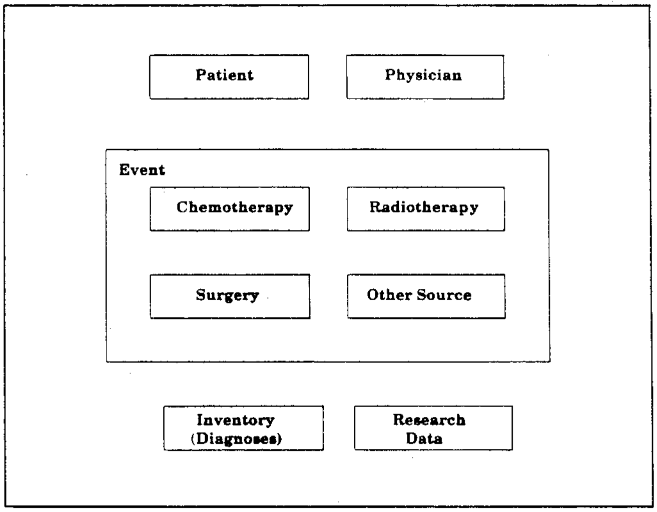


Fig. 1. Major subsystems of the NOIS database

The Neuro-Oncology Information System

We have designed and implemented an information system that addresses the data problems of collaborative researchers and meets all of the design specifications stemming from an analysis of those problems.

Major elements of the database

The major elements of the NOIS database are shown in Fig. 1. Patient and Physician subsystems are kept because they are required for patient follow-up. The treatment subsystems are modeled as "events" in NOIS and consist of surgery, chemotherapy, radiotherapy, and other, where the "other" subsystem stores follow-up data that are not directly linked to any of the treatment modes.

A tissue-inventory subsystem stores the various tumor diagnoses, as well as a detailed hierarchy of dissection, distribution, or current location of the tissue.

Some researchers have chosen to keep their private research datasets in NOIS, and we have a subsystem that accommodates this need. These datasets are accessibly only to users who are explicitly named by the researcher.

The "Case" concept

"Case" is a unifying concept in NOIS. Figure 2 shows it in its role at the center of the other subsystem. A NOIS Case can be thought of as a hook onto which all other data about one surgery for one patient can be attached.

NOIS allows for the data in subsystems emanating from Case to be input in any order. The input process also provides linking mechanisms for all the various data to be linked to the same Case. One drawback of this design is that clinical and inventory data for the same Case may be

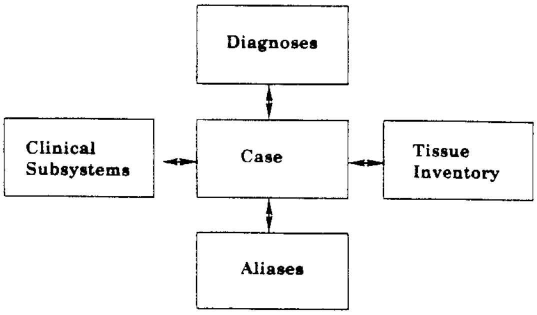
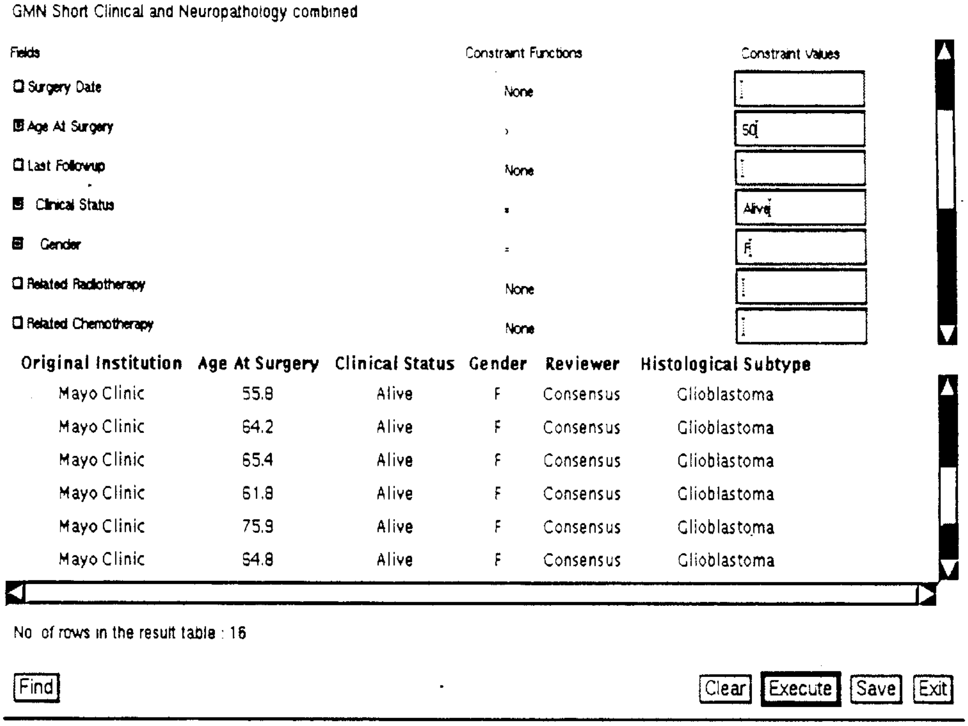


Fig. 2. The role of Case in NOIS

Fig. 3. A sample query in the NOIS Browser: data analysis



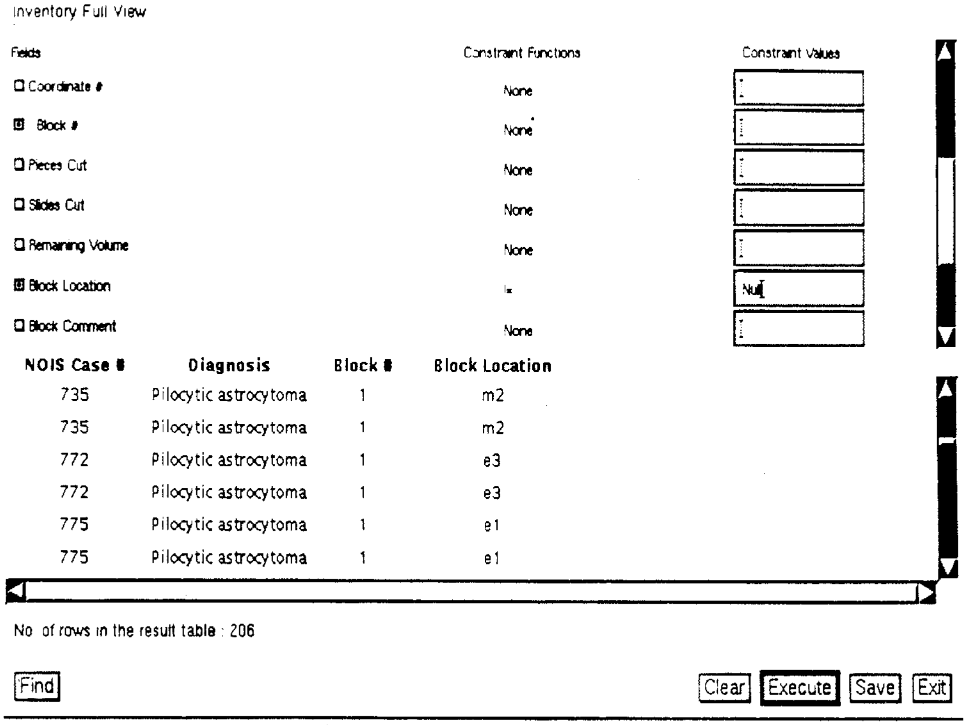
GMN Short Clinical and Neuropathology combined

Fields	Constraint Functions	Constraint Values
□ Surgery Date	None	
☒ Age At Surgery	>	50
□ Last Followup	None	
☒ Clinical Status	=	Alive
☒ Gender	=	F
□ Related Radiotherapy	None	
□ Related Chemotherapy	None	

Original Institution	Age At Surgery	Clinical Status	Gender	Reviewer	Histological Subtype
Mayo Clinic	55.8	Alive	F	Consensus	Glioblastoma
Mayo Clinic	64.2	Alive	F	Consensus	Glioblastoma
Mayo Clinic	65.4	Alive	F	Consensus	Glioblastoma
Mayo Clinic	61.8	Alive	F	Consensus	Glioblastoma
Mayo Clinic	75.9	Alive	F	Consensus	Glioblastoma
Mayo Clinic	64.8	Alive	F	Consensus	Glioblastoma

No of rows in the result table : 16

Fig. 4. A sample query in a NOIS Browser: inventory



Inventory Full View

Fields	Constraint Functions	Constraint Values
□ Coordinate #	None	
☒ Block #	None	
□ Pieces Cut	None	
□ Slides Cut	None	
□ Remaining Volume	None	
☒ Block Location	!=	Nul
□ Block Comment	None	

NOIS Case #	Diagnosis	Block #	Block Location
735	Pilocytic astrocytoma	1	m2
735	Pilocytic astrocytoma	1	m2
772	Pilocytic astrocytoma	1	e3
772	Pilocytic astrocytoma	1	e3
775	Pilocytic astrocytoma	1	e1
775	Pilocytic astrocytoma	1	e1

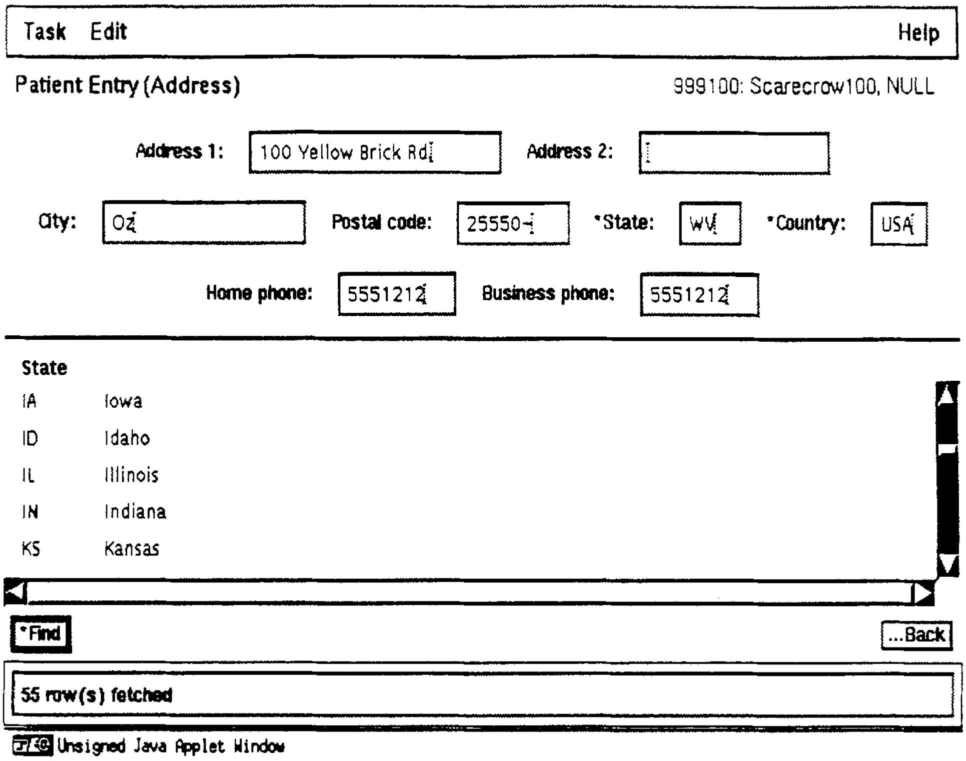
No of rows in the result table : 206

entered but never linked. We believe that the advantage of reflecting the real input process, whereby it is simply not possible to order the input, outweighs this disadvantage. Also, in our monthly data reviews (see below), we check on the status of unlinked subsystems and try to keep them linked properly.

Data-quality issues

We have a three-pronged effort in place to maintain high-quality data in the NOIS. First, some data are constrained at data-input time to conform to certain allowable values. For example, the surgery date is not allowed to occur before the

Fig. 5. A sample screen in a NOIS input form



Patient Entry (Address) 999100: Scarecrow100, NULL

Address 1: 100 Yellow Brick Rd
Address 2:
City: Oz
Postal code: 25550-
*State: WV
*Country: USA
Home phone: 5551212
Business phone: 5551212

State	
IA	Iowa
ID	Idaho
IL	Illinois
IN	Indiana
KS	Kansas

55 row(s) fetched

birth date. Similarly, for some histology subtypes that comprise a certain diagnosis, the grades are constrained to certain values.

Second, whenever possible, we disallow free-form data input. For all fields whose value sets are predefined, we allow input only from a choice of lookup values that we maintain in the database. This prevents a good deal of erroneous data input due to misspelling, capitalization, or punctuation. One example of an especially useful lookup table is the official World Health Organization numbers and names for every possible histology subtype for brain tumors.[3]

Our third effort is regularly scheduled monthly data-quality meetings in which we review data values, looking for inconsistencies or incompleteness. When we discover problems, we correct them and, when possible, implement constraints that will disallow or warn about these errors at data input time.

Data privacy

Patient privacy is protected in the NOIS database by way of a *role* mechanism. Appropriate permissions on database objects, such as tables and views, are granted to roles. Roles are then granted to individual users as appropriate. For example, in NOIS we have created a role called "clinical input," and we grant this role only to those users who are designated to input data into the clinical subsystems in the database. Another role, "inventory input," is granted only to those users who are allowed to update data in the inventory subsystem.

Researchers are not granted any roles that would allow them to see patient-identifying information. They retrieve data on the basis of anonymous patient or tumor identifiers.

Access to NOIS

The NOIS is implemented using a client-server architecture. In this model, the database[4,5] plays the role of the server, and all users are viewed as its clients. Clients input and query data in the NOIS database using forms accessible through their local World Wide Web browser.[6,7] The NOIS home page is located at URL http://acuity.cis.ohio-state.edu.

A sample screen from the NOIS Browser is shown in Fig. 3. This query might be formulated and executed by the user in the data-analysis portion of the research. The screen is divided into two areas, the top section, in which the user builds a query, and the bottom portion, which holds the results of a query after it has been executed. In the example, the user has asked to retrieve all records for female patients from the Mayo Clinic whose age at surgery was greater than 50 years, who are still alive, and whose consensus diagnosis was Glioblastoma.

Another sample screen from the NOIS Browser is shown in Fig. 4. This query might be executed while researchers are designing their experiments. In this query, the user is asking to see all NOIS cases in which the histology subtype is Pilocytic Astrocytoma and there is tissue in the freezer. For this query, 206 records were returned, indicating that there were 206 frozen blocks.

A sample screen from a NOIS Input Form is shown in Fig. 5. Patient-identifying information has been disguised. The context in the upper right-hand corner of the screen shows that this is a clinical-information screen for a patient with hospital identifier 999100 and with name NULL Scarecrow 100. In this screen the user has learned that the home state of the patient is not West Virginia, and she wishes to change it to the correct state, say Indiana. She has used the Find button in the lower left-hand corner of the screen to call up the allowable entries for the State field, and these are shown in the lower half of the screen. When the user clicks on any one of these, it will become the new value of the State field.

Client software is implemented using Java, a programming language that supports two important features.[8,9] The first Java feature is that client software is written as applets (i.e., little applications) that are downloaded to the client platform when the NOIS forms are invoked. This means that only one version of the forms is required and eliminates the version-control problems that often plague clients.

A second feature of Java is that it is platform-independent. The Java compiler creates a byte code that is interpreted on each different platform to have identical behavior. This means that clients can use PCs, Macs, or workstations running Unix and still be able to access NOIS data easily.

A third desirable characteristic of Java is that all the medical institutions in the Glioma Marker Network allow the downloading of Java applets to go through their firewalls. Thus, we have not had to obtain any special permissions or exceptions for our researchers to access the database.

Tissue aliasing

When tumor tissue is moved among departments in a given research institute, and especially when it is shared among collaborating institutions, it is usually given an identifier in each department and at each institution. For example, tissue may be given a neuropathology number in the home Neuropathology Department that manages the tissue, say NP 3478. If a researcher in the home institution chooses to use a piece of that tissue for a local experiment, the researcher typically assigns another identifier, say YATES-GMN 260. If the tissue is shared with another institution, it may receive yet another identifier, say UNC-GMN MS 954328.

Because all of these pieces of tissue are from the same case, and because it is convenient for researchers and others to assign additional identifiers to tissue, the NOIS provides an aliasing mechanism to accommodate and support this practice. Tumor tissue from each case may have any number of aliases, and researchers may refer to that tissue by any one of them.

Requirements of NOIS users

A database is only as good as the quality of the data that are stored in it, so collaborating researchers must themselves commit resources to input data in a timely way and to maintain the data in as complete and correct a manner as possible. Typically, at least one staff member at each collaborating institution must be assigned a new role to accommodate the data needs of the collaborative research.

New role for local staff

Among the responsibilities of the researchers is to assign unique, non-patient-identifying identifiers for each patient and for each tumor in the collaborative studies. The numbering task is a complex, difficult problem among laboratories and departments in a single institution, and is even more complex and difficult among collaborating institutions.

In our experience, we have seen problems stemming from the use of variable formats for the same identifier. For example, the same patient was identified as 93042, 093042, 93-04-2, and 9304. Problems have arisen because the pathologists reversed the patient and tumor numbers on their pathology forms. Problems have also arisen because patient identifiers from different aliases were combined ambiguously. For example, on one list of patient/tumor numbers, about two-thirds of the numbers were for a YATES-GMN alias, but the other third were taken from the unique NOIS number space.

Ideally, the goal is to catch these problems at the data-input stage and avoid putting incorrect data into the database. When such erroneous data are input, quality-control meetings may be used to try to track them down. In either case, a careful and deliberate interaction between the institution and those managing the data is an absolute requirement.

Data input

Collaborating institutions must provide clinical data for their local patients who become a part of the collaborative research. In the NOIS, we provide three options for such data input.

The first option is hardcopy input. We have a hardcopy clinical-input form, and the institution simply fills in the form and sends it to the NOIS maintainers. We have created an intermediate electronic form in which to store these data and automated scripts that read the intermediate form and upload the data correctly into the NOIS data tables. An error log is created when necessary. This option is the easiest for users, because no training or other programming is required.

A second option is to provide the data by way of an electronic file transfer where the data are stored in a prespecified electronic form. Institutions choosing the data-input mode have programmed scripts that read their internal clinical-data formats and translate them into the newly prespecified electronic format. We in turn move this into our intermediate electronic form and then run the same automated scripts as for the hardcopy to upload the data into NOIS. This option is ideal for institutions that already have all the relevant clinical data electronically coded and have the extra programming resources to write the translation program.

A third option is to provide data by way of our Web-based clinical input forms. This option is the most interactive, in that the forms constrain the order and some values of the data. If there are errors in the data format, the user will be notified immediately and asked to correct them. This option is the most difficult for most institutions, because the use of the forms requires training, and remote training is often difficult to do.

Lifelong support

Any software system, including an information system, requires birth, intensive early nurturing, and ongoing long-term maintenance. This is true because requirements change over time, as does computer technology.

In the NOIS, the project has been transformed from its beginning goal of supporting clinicians at one institution, to supporting clinicians at multiple institutions, to its current mission of supporting researchers at multiple institutions. Each of these transformations required extensive database and form changes.

Since the beginning of the NOIS project, computing technology has changed dramatically. In addition to dealing with the continuing, evolutionary upgrades of software from vendors whose applications we are using, we have had to deal with the revolutionary impact of the Java programming paradigm.

These continuing changes must be understood and supported by researchers. An information system to support collaborative research is never "done," because at any given time something is being improved, be it the database, the input or query forms, the quality of the data, or the hardware/software platform on which the information system depends.

Conclusion

For research centers and interinstitutional groups studying human brain tumors, it is essential to have a computerized information system. Such a system must allow the investigators easy access to enter and retrieve clinical and research data at the same time as assuring anonymity of the patients and security of the data.

The NOIS that we have developed at The Ohio State University has these characteristics. It is fully operational and is being accessed remotely by all reseachers and their staff in the Glioma Marker Network at The Ohio State Univeristy, Johns Hopkins, Mayo Clinic, the University of North Carolina, the University of California at San Diego, and the University of California at San Francisco. The NOIS currently contains clinical, diagnostic, tissue-inventory, and research data. Approximately 1800 patients have been registered in the NOIS. Close to 800 cases have consensus diagnosis. The number of cases for which we store both clinical data and consensus diagnoses is nearly 700. Detailed tissue inventory is kept only for The Ohio State University inventory, with tissue recorded for close to 450 cases, but the NOIS has the ability to track tissues at the other five institutions.

Acknowledgments We would like to acknowledge the contributions to the NOIS project of our many colleagues. Among the many graduate students in the Department of Computer and Information Science who worked on the project, we acknowledge the contributions of K. Bowers, J. Boyd, S. Feitler, M. Geroch, M. Hu, T. Hudek, S. Hussain, S. Jennings, J. Loomis, R. Mazuk, K. McGrory, J. Miller, I. Ordonez-Rienoso, G. Patel, K. Patel, D. Pati, R. Rao, N. Reams, G. Schueman, S. Sinha, S. Subramanian, S. Vellimedu, A. Wang J. Zapatosky, and A. Zwerin. Dr. Sandra Cottingham, formerly in the Division of Neuropathology at The Ohio State University, assisted us in understanding the tissue inventory process.

Supported in part by grant 5 P20 NS31087-03 from the National Institute for Neurological Disease and Stroke, and in part by grant 5 U01 CA50910-06 from the National Cancer Institute.

References

1. Department of Computer and Information Science, The Ohio State University (1996) Neuro-oncology information system technical reference, June 1996. Also found at http://acuity.cis.ohio-state.edu
2. Mamrak SA, Boyd J, Ordonez I (1997) Building an information system for collaborative researchers. Software Practice and Experience, Vol 27, No 3, March 1997, pp 253–263
3. Kleihues P, Burger PC, Scheithauer BW (eds) (1993) Histological typing of tumours of the central nervous system, 2nd edn. Springer-Verlag, Berlin The World Health Organization international histological classification of tumours
4. Chen PPS (1976) The entity-relationship model – toward a unified view of data. ACM Transactions on Database Systems, Vol 1, March 1976, pp 9–36
5. Codd EF (1970) A relational model for large shared data banks. Communications of the ACM, Vol 13, No 6, June 1970, pp 377–387
6. Bowers SK, Sinha S, Mazuk RE, et al. (1997) A data browser for the neuro-oncology information system. Technical Report, The Ohio State University, July 1997, Internal NOIS Report
7. Sinha S, Bowers SK, Mazuk RE, et al. (1997) Storage and retrieval of medical research data on the World Wide Web. World Wide Web (submitted for publication)
8. Friedel DH, Potts A (1996) Java programming language handbook. Coriolis Group Books
9. Jaworski J (1996) Java developer's guide. Sams Net