Confidentiality, Data Security and Cancer Research:

Report of a Workshop

INTRODUCTION

Medical research frequently requires the collection, storing, transmission, sharing, and analysis of sensitive information about the health or behavioral characteristics of individuals. Participants in research expect that data identifying them will be kept private and protected from unauthorized use. The research enterprise requires, therefore, that a fine balance be struck between the need to protect sensitive information that identifies individuals and the need to provide researchers with access to the information in a manner that will permit them to do their work. Both of these serve important societal goals.

Effectively protecting the confidentiality of individually identifiable data requires that comprehensive practices be implemented broadly. The biomedical research community has a generally successful record of safeguarding the confidentiality of individuals participating in research. There is no reason to think that a systemic problem with data confidentiality presently exists within the research enterprise. On the other hand, striking advances in biomedical science and the proliferation of electronic data storage, linkage, and transmission have created significant new challenges in maintaining confidentiality. The goal of the research community, and of society generally, should be to protect the confidentiality of research participants without compromising the critical research necessary to improve human health.

NATURE OF THE RECOMMENDATIONS

The recommendations summarized here represent a consensus of Workshop attendees. The participants strove to develop recommendations which, if implemented, would provide a high level of data security and, at the same time, not impede the research process unnecessarily. The participants agreed that absolute data security was not a realistic goal, and attempts to provide it would seriously obstruct the conduct of research. Achieving the appropriate balance between high levels of data security and continuing access to information is a delicate process. It should surprise no one that the Workshop attendees, coming from a variety of backgrounds, did not agree in all details on how this balance should be achieved. Despite this, there was a remarkable degree of consensus on the recommendations.

This document describes, in effect, a set of “Good Security Practices” (GSP) and should serve a function analogous to the standards for Good Manufacturing Practices (GMP) that inform the production of pharmaceuticals. Good manufacturing practices serve to insure that the production, testing, and handling of compounds intended for human use conform to high standards and that the compounds themselves satisfy certain criteria for purity, safety, stability, and bioavailability. They do not guarantee that drugs so produced will have a trouble-free sojourn in the clinic, but adherence to them makes the probability of avoidable difficulties much less likely. The cost to pharmaceutical sponsors in terms of procedural detail is considerable, but the general welfare and, in the end, the interests of drug sponsors trump expediency.

It is always possible to argue that the recommendations in this document do not go far enough. Institutions and organizations that wish to implement more stringent measures will have to

consider carefully whether the additional procedural price for more stringent security protections is worth the additional security that is achieved. It is certainly in society's interest that the research community not impose on itself, or have imposed on it, more stringent measures than are, in fact, necessary to assure an acceptably low probability of security violations.

FORMAT OF THE RECOMMENDATIONS

Despite the diversity of research that formed the context of these discussions, Workshop participants were able to define a large zone of commonality among the areas. The recommendations are organized in a manner that discusses these common areas in general terms, without explicit reference to minor differences among the various areas of research. For example, how a research organization should handle paper documents bearing individually identifiable information does not depend on whether the information deals with genetics, behavior, epidemiology, surveillance, or clinical trials. In a number of instances, however, the specific research context turned out to be an important determinant of the specific recommendations, and these are noted.

DEFINITIONS

What kind of information can lead to identification of an individual? Unfortunately, this question has no simple answer. Name, address, and social security number probably satisfy anyone's definition. Consider, however, the case of individuals with certain unusual diseases. Such people – rare and perhaps conspicuous for one reason or another - may be well known in their local communities. It might require very little information to violate their confidentiality. Quite innocent information like age, sex, and hair color might link potentially sensitive medical details with them unambiguously. In a small community, even release of a pedigree might lead to identification of individuals without actually naming anyone.

These examples show that the term "individual identifier" does not lend itself to absolute definitions. Certain pieces of information about people might or might not serve as identifiers, depending on the time, energy, and access to other information that one brings to the search. As if in tacit acknowledgment of this, the Health Insurance Portability and Accountability Act (HIPAA) legislation defines individually identifiable health information as health information that "...identifies the individual; or with respect to which there is a reasonable basis to believe that the information can be used to identify the individual." These elements include, but may not be limited to, name, address, zip code, names of relatives or employer, birth date, telephone or fax number, e-mail address, social security number, medical record number, health-plan beneficiary number, certificate or license number, vehicle serial number, web URL, Internet Protocol (IP) address, finger or voice prints, photo images, or any other unique identifying designation that may be available to any possessor of the information. The more that these items, or other potentially identifying personal data, are removed from information or specimens, the less likely it is that individuals can be identified from the information that remains.

In this summary we use the term "anonymize" for specimens or data from which all identifying information has been stripped, and all links to the source permanently severed, so that there is no

reasonable basis to believe that an individual can be identified. Such specimens or data cannot any longer be associated with individuals, and any scientific information derived from them in research cannot find its way back to the health records of the individuals from whom the tissue or data were derived. We use "de-identified" to indicate the deletion of information from tissue samples or from research data that might reasonably lead to the identification of individuals. Information or specimens that have been de-identified in this way may retain coded or encrypted links with identifiable information in source records or repositories. Retention of these links permits the flow of information, either unidirectionally (from source to research record) or bidirectionally, if the individual holding the key or code permits this to happen. The aim of de-identification is to make the probability of individual identification by those using data in research remote.

Throughout this document we distinguish "research data" from general health-related information. Certain kinds of medical data, however, cannot be unambiguously dichotomized in this way. Obvious examples are the routine radiographs and blood tests specified by a protocol to assess response to therapy in a clinical treatment trial. Such data ordinarily become part of both the medical record and the research record. By contrast, research data that have not yet been validated for use in the health-care setting and for medical decision-making do not generally become part of an individual's health and medical record. Exceptions to this generalization may occur when data stemming from the research have clear implications for patient care (for example, plasma levels of an investigational drug if it is already known that levels correlate with the probability of toxicity).

EDUCATION AND MONITORING

Protecting the confidentiality of identifiable research information requires the development and implementation of education programs and the formal credentialing of staff; the formal designation of an individual to assure compliance with established institutional standards; and the establishment of institutional policies and procedures to deal with breaches of confidentiality.

Institutions and other groups conducting research are responsible for developing specific, explicit procedures for handling data, including physical and electronic data transport. Every organization should require confidentiality training for all employees who will have access to patient specimens or data and should require employees to sign a statement of compliance with the procedures on a scheduled basis. All levels of staff must participate, including physicians, scientists, nurses, messengers, clerks, secretaries, abstractors, and information technology personnel. Contractors who handle data should also be included in the training process and security procedures. Training and re-credentialing should be on a regular basis, much as training for Universal Precautions is currently implemented in hospitals.

Small physician offices that conduct clinical research should participate in confidentiality training for all staff and should have in place confidentiality procedures consistent with GSP. The sponsoring organization or research base with which the office is affiliated, such as a multi-

center research consortium or a cancer center, should facilitate this training by providing affiliates access to their training programs. The clinical research community should develop training materials targeted to the special circumstances of researchers who operate outside an institutional setting.

The educational program should include information about the lack of security inherent in faxing, e-mailing, and other electronic data transfer. It should include reminders about not using names or other personal identifiers in conversations in public areas such as open labs, elevators, or hallways. It should remind employees of the special duty to maintain confidentiality when research involves individuals they know personally or public figures.

In addition to confidentiality manuals, the research community should develop and use web-based instruction programs and videos on the basic principles and issues of research confidentiality, and the potential for harm from a breach of confidentiality. Local institutions may then supplement these general materials with specifics relevant to their own procedures and special circumstances.

Entities sponsoring or participating in human subjects research are responsible for educating their local communities regarding the strategies they have implemented to protect the confidentiality of research data. The goal should be to diminish any misperceptions that may exist regarding the confidentiality risks associated with participating in research. These education programs should be responsive to local community needs and concerns.

Adherence to institutional standards will best be assured through the establishment of "confidentiality compliance officers" specifically designated within each research facility. Alternatively, if institutional mechanisms already exist to oversee general data security, these may be expanded to include oversight of research data. The responsible individual(s) would oversee educational programs and adherence to them, collect signed confidentiality statements, and take or recommend remedial action in the case of a breach of procedures. Methods should be in place for actively monitoring implementation of procedures to insure confidentiality, rather than for simply reacting to breaches. Each institution should establish a remediation process for individuals who breach standard procedures. Institutional incentives for good security programs might be offered.

BREACHES OF CONFIDENTIALITY

All research data should be handled in such a way as to keep the risk of breaches in confidentiality and privacy to an absolute minimum.

Although breaches can occur at any level, ultimately, the principal investigator of the study is responsible for seeing to it that the probability of a breach of confidentiality is minimized. Investigators are responsible for notifying the IRB and other responsible parties when a breach of confidentiality is suspected or has occurred. Educational materials should include information regarding whom to notify about possible breaches of confidentiality. The contact individual might be the confidentiality compliance officer within the organization or an ombudsman.

Policies and procedures for dealing with breaches of confidentiality should be explicit, rather than implied. Penalties for misuse of identifiable data and breach of confidentiality must be included in institutional policies and procedures.

When a breach of privacy or confidentiality is alleged, there should be a procedure in place for a valid, objective, adequate assessment of the facts, such as an investigation by outside counsel. The outside counsel would evaluate the allegation and ascertain the facts. If the investigation finds evidence of wrongdoing or sufficient cause, the matter should be referred to the Human Resources Department, the IRB, or other appropriate institutional committee for adjudication. If the IRB or other committee concurs that significant wrongdoing may have occurred, the case should be referred to the appropriate institutional committee, such as the faculty senate or the medical practice committee. Obviously, the physician of record should also be notified.

The IRB should evaluate the need for notification of participants if a breach concerning them is discovered. The involved research participants should then be informed about the outcome of the fact-finding process, any actions that have been taken, and any recourse they may have.

A breach of confidentiality or privacy is an act that warrants significant penalties. Specific sanctions should depend on the severity of the violation. They may include employee disciplinary action and any of the following: remedial training in confidentiality, loss of certification of competency in confidentiality (local only), prohibition by the IRB of future participation in research at the institution, dismissal, and notifying the appropriate professional licensing boards of the findings. If the violation is severe as judged by an appropriate institutional body, the funding agency or sponsor should also be notified.

EQUAL PROTECTION FOR ALL INDIVIDUALLY IDENTIFIABLE INFORMATION

The collection, processing and storage of identifiable information on individuals should treat all classes of information equally. The degree of protection should not depend on the type of information: an individual's name, birth date, and social security number should be afforded the same degree of protection. Similarly, all specimens and individually identifiable data associated with the specimens should be treated as equally sensitive. If data are specific enough to allow identification of an individual, the data must be protected by applying rigorous rules of access that include justification for the need to know.

Within a given study, all individually identifiable research data should be handled in the same way. Any error should be made in over-protecting all data, not in establishing a minimum standard that may be insufficient for individually identifiable data. The concern that genetic information is more sensitive than other classes of sensitive health-related information is not sufficiently compelling to warrant different levels of protection, although genetic information on an individual may affect other family members.

Surveillance data bases merit special consideration. Collection of these data is authorized, and sometimes required, by statute; from these databases public-use files are often made freely available with no restriction other than prohibition of intentionally identifying individuals. These

files are created with the intention of minimizing the chances of individual identification. Clearly, the more data that are made available, the greater the chances that cross-tabulations according to many specified characteristics may produce small enough cells to identify individuals under some circumstances. For this reason, curators of surveillance databases should take measures to minimize the chances of identification of individuals as the data resource is being used. This is a particular issue when data files are released for research that contain more data elements than are included on public use files, or that contain frankly identifiable data. Data that make it easier to identify individuals are often just those data elements that are particularly valuable for surveillance research and that tend to insure accurate consolidation of information from multiple sources into maximally informative files. Under these circumstances a formal review process should be in place that oversees the access of researchers to the data.

There are a number of approaches to minimizing the chance that surveillance data files will result in the inadvertent identification of individuals. Among these are:

- Tailoring release of data to the requirements of the particular researcher and not releasing more data than the researcher needs to answer the questions posed by the proposed study; identifiable data and other data elements that increase the potential for identifying individuals should be released on the basis of what the particular researcher needs;

- Fragmenting released data into multiple, non-overlapping sets of files, to make extensive cross-tabulation and the generation of small cells impossible;

- Reviewing in real time the tables generated by researchers when querying large databases, to assure that inappropriately small cells are not, in fact, generated by the data queries.

Whichever method is used, it is prudent to treat derived files with the confidentiality protections accorded to potentially identifiable data.

SAFEGUARDING IDENTIFIABLE INFORMATION

Institutions and research organizations should develop policies and procedures to assure the security of both paper and electronic records containing identifiable information. For certain classes of data, such as cancer surveillance information, established policies and procedures for data storage and transmission have existed for some time at participating institutions. As new information technologies become more widely available, these policies and practices must be revised and updated to assure continued effectiveness. For other classes of data, such as cancer epidemiology, behavioral sciences, and clinical trials, confidentiality policies and practices may be relatively new to most research organizations, and their formulation and implementation will require substantial effort and resource investment.

Whatever the research area, data storage and transmission policies and procedures must be in place in all institutions. These include technical specifications of how data are processed, stored, and transmitted physically and electronically; who is responsible for assuring protection of the

data; the use of encryption and other protective measures; and methods to monitor adherence to and effectiveness of such policies and procedures. As part of this, organizations should also consider the need for a "vulnerability assessment" that would serve to raise institutional awareness of local confidentiality issues and address the adequacy of the security measures that are in place. Ongoing assessments of electronic systems should include the periodic running of diagnostics to assess the integrity of the system.

Two important legal safeguards that help to insure confidentiality are the Privacy Act, which covers data in the possession of the federal government and Certificates of Confidentiality (http://cancertrials.nci.nih.gov/researchers/safeguards/certificates/index.html), which are available for federally funded projects. A national umbrella Certificate of Confidentiality for all genetic studies is under consideration at this time because of public concern about access to genetic data.

ELECTRONIC RECORD SECURITY

A recent Institute of Medicine publication (For the Record - Protecting Electronic Health Information, National Research Council, National Academy Press, 1997) contains a set of excellent recommendations. The Checklist of Security Measures Recommended for Immediate Implementation includes authentication of users, access controls, audit trails, disaster recovery, remote access procedures, external communications, software discipline, system assessment, policies and procedures for security and confidentiality, an information security officer, education and training, sanctions for violations, authorization forms, and patient access logs. We comment on several of these items below. Note that data collected as part of international collaborations should be subject to the same protection measures.

Institutions and coordinating centers should have the following *technical practices* in place:

- Authentication of users by means of passwords or digital ID;
- Access control by means of role-based authentication/access, locked server room, and an internal firewall;
- An audit trail that documents who, when, and for what purpose data were accessed;
    - A disaster prevention and recovery plan including adequate fire and entry alarms where data are stored, a fireproof file space for paper, routine backups of electronic data at intervals appropriate for the rate of data accrual;
- External firewalls in places of remote access;
- Software discipline should include virus checking, the use of the minimum appropriate software with no extraneous software (i.e., games which could have impact on data);
- System assessment including diagnostics runs and external audits should be conducted regularly to insure the integrity of the system and absence of holes.

The *transfer of data* should be accompanied by:

- A data-transfer agreement incorporating confidentiality standards to ensure data security at the recipient site and set standards for data use at the recipient site;

- A paste (electronic) or stamp (paper) on all records containing identifiable data as a reminder of the need for special handling.

The *maintenance of confidentiality of electronic records at local sites* should include:

- Use of codes or encryption to separate patient identifiers from research data;

- Designation of an individual custodian of coded or encrypted identifiers, which should be maintained separate from research data;

- Release of identifiers on a need-to-know basis only.

For certain types of data repositories, it is not practical to prohibit storage of identifiers. For example, the data coordinating centers for clinical trials may, in fact, be the only source of continuity for long-term follow-up, since care-givers and the locus of care for particular patient participants in research may change over time. For such repositories, additional measures are necessary:

- When it is necessary to maintain the link with patient identity, such as in a large clinical trial with long-term endpoints, such items as name, social security number, and other obvious identifiers should not be used routinely in analyses and should not be stored with the information as secondary identifiers.

- Obvious identifiers should be retained in a secure area/electronic file where only a few authorized individuals have the ability to link them back to the full data set.

- Name should not be included on documents sent from clinical centers; a unique study number and patient initials at most may be used as a cross-check. For example, the pharmaceutical industry standard is never to obtain name or social-security number. Verification of informed consent is accomplished during visits to the local site, and not transmitted to the central office.

Additional Comments:

- Telecommuting and the use of home offices are likely to increase in the future and will create new challenges for data security. Off-site use of individually identifiable data requires the same level of security and procedures to address special issues, including data-transfer agreements, secure transmission procedures, and encryption. Additional safeguards are needed as well, including: maintenance of minimal data on home computers, use of electronic screen savers, and password control at home. When it is

anticipated that confidential data will be present in an investigator's home, the nature of the data, the restricted circumstances under which home use and storage is appropriate, and home-specific confidentiality procedures should be clearly specified in writing for each study or program.

- Other recommended safeguards for electronic records include: segregation of the research record from the medical record and the use of restricted views employing field-level security and role-based authentication. In the future, access to data will probably be afforded less by physical data distribution and more by restricted views of a portion of a data base. Then, as now, responsibility for data security and integrity belongs to all the individuals who are in actual possession of the data.

PAPER RECORD SECURITY AT LOCAL SITES

The maintenance of confidentiality for paper records should include:

- Restricting access to data-storage areas, the use of locked file rooms or cabinets in limited-access areas, a forms tracking log for any external disclosures, and a sign-out system for internal use of data.

- Restricting access to the databases themselves. For data repositories to which regular and frequent access by a core set of authorized personnel is necessary, a global statement of standing access by these individuals is sufficient and obviates the need for recording each access event. For authorized individuals outside this core set who need occasional access to data for their research, an audit trail should be in place that documents who, when, and for what purpose data were accessed.

- Developing and implementing institutional policies for the secure transport of information from one physical location to another.

- Assuring confidentiality of written evidence that a patient is on a specific research study; for example, logs or lists of screened individuals or participants should not be left out on desks or in other open-access areas.

- Safeguarding of ancillary records, e.g., pharmacy records or data on patients screened for clinical trials participation.

- Situating FAX machines in secure or limited-access areas; use of pre-coded phone numbers to eliminate dialing errors; cover sheets so data are not physically exposed; testing FAX machines to insure correct number and function; and de-programming FAX memory storage after use to prevent recovery of confidential information.

- Employing established shredding procedures for disposal of documents after use.

SECURITY FOR HUMAN SPECIMENS AND SPECIMEN REPOSITORIES

- Networks and centers that maintain collections of human specimens for use by researchers not connected to patient care should establish adequate policies and procedures to ensure that researchers do not have access to personally identifiable information relating to the individuals from whom the tissues were derived. The pathologist or other individual who submits specimens can act as an "honest broker" by removing personal data prior to submission of the specimen and data to the resource. For example, the NCI-supported Cooperative Human Tissue Network deletes all personal identifiers from the pathology reports it distributes with tissue samples. The Cooperative Breast Cancer Tissue Resource assigns unique coded identifiers to the data submitted to the central database, which does not have a link to the patient record; only the code is attached to specimens when they are sent to researchers. Other institutions encrypt personal information before the sample is sent to the repository.

- Any institution that serves as a source of tissue and data for research should have the same confidentiality standards and procedures for its own internal research collections as formal specimen repositories have for theirs. These standards and procedures for research collections should apply to individual pathologists as well as to resource managers. When a pathologist plans a research project using specimens from the archive (s)he maintains, another pathologist should serve as the "honest broker" to remove identifiers from the specimens to be used in the project.

- As for any sensitive records, repositories should have codes or encryption routines, passwords, and other computer security features in place to ensure against violations of confidentiality. Other computer security features, such as lookup logs or records of standing access, should also be used to track access and prevent unauthorized access to clinical information. All hard copy files should be stored and handled in a secure manner, with the protections described previously for paper records.

- In research laboratories biologic samples should be identified by a code only, not by name. Personnel in research laboratories should not be able to link the code to a specific name. Lines of accountability in releasing information should be clearly defined.

ADDITIONAL DATA SECURITY CONSIDERATIONS FOR EPIDEMIOLOGY

For IRB-approved research, entities that are responsible for individually identifiable health data may release de-identified data for which the recipient of the information cannot link the data back to an individual, even if a key exists somewhere. Linkage may be possible in theory, but prohibitive in practice. In some situations, the best use of data for epidemiologic and behavioral research requires that certain critical variables be included in the data set. Therefore, the best practice is to have a system where all reasonable attempts are made to de-identify individuals, and where penalties are established for those who attempt to breach confidentiality. Reasonable attempts may include statistical manipulations of the data to obscure individual identities, as well as the removal of specific identifiers not required for the conduct of the research.

Entities that are responsible for research data should designate an individual to hold the key or code that allows for re-linking of data sets to the original health-care files that contain patient

identifiers; this individual might be the confidentiality compliance officer or his/her designee. Data should routinely flow only one way: from the primary health-care files to the data recipient. This prevents newly acquired research information that has not yet been validated for medical decision-making from reaching the primary health-care files of the study participant. The one-way data flow is not intended to restrict an epidemiologist or other researcher from reporting back to a registry updates of routine data, such as changes of address or deaths. There may also be circumstances in which data generated in the course of an investigation have a high likelihood of benefitting the research participant, if the data were available to the participant and his/her caregivers. Examples might include an unanticipated positive test for a communicable disease or a major change in diagnosis upon review of pathology sections. Approval for placing such information in the health-care record and notifying the participant or care-givers should be by an IRB or institutional bioethics committee.

The protections afforded by a Certificate of Confidentiality are attractive for some epidemiologic studies. Protection with such Certificates extends both to participants and providers, and disclosure cannot be compelled by subpoena.

HANDLING LINKAGE OF IDENTIFIABLE INFORMATION

Principles protecting the confidentiality of files used for linkages are largely identical to general principles for storage, release, and use of any identifiable data. The unique aspect of linkage is that it typically involves two or more files held by different parties; each file contains personal identifiers, and linkage of two or more files allows identification of individuals in each file with attributes not originally present in that file. Written procedures must be developed to specify which party will release, or re-release, its file to the other party, and how the party receiving the linked information will ensure protection of this information in its storage and use.

The IRB should review and, where necessary, approve all procedures for the collection, storage, coding, encryption, and rules for release of specimens and associated linked data for research purposes. The linkage procedures themselves should be developed by a group with appropriate expertise; the specific group or committee will vary according to the institution. For multi-center studies, a central review mechanism may be appropriate for ensuring consistent application of the linkage of archived information.

In behavioral research and epidemiology the linking of information from multiple data sets is an essential activity. Support of this practice should be insured in procedures developed to protect the confidentiality of individually identifiable data. Such a capability is possible if procedures of de-identifying and coding data are accompanied by penalties for attempts for trying to re-identify data. Here again, the responsibility for adherence lies with the recipient. This can be accomplished through a contractual agreement and IRB approval.

In genetics studies, linkage of archived information can involve the creation of a pedigree. Each pedigree illustrates a unique family that can potentially be identified. Debate continues about how best to present pedigrees in publications to minimize the likelihood of identifying the family, without distorting information important to the interpretation of the results of the study.

Falsifying pedigrees, as some have suggested, is not an acceptable solution to this problem.

DE-IDENTIFYING AND ANONYMIZING

In most research contexts, anonymizing data is simply not possible; the maintenance of identifiers is necessary to realize the scientific goals of most studies in all categories of cancer research. In a clinical trial, for example, any study requiring continued follow-up to determine outcome requires maintenance of a link to the source of the participant's data. Certain secondary uses of data from the same trial, however, may permit use of anonymized information, if the need for long-term follow-up is not an issue.

Similarly, in many epidemiologic, behavioral, or genetic studies, it is virtually impossible to anonymize data completely when many variables are considered or when a number of genetic markers are genotyped. Anonymization, however, is not really the issue. The primary concern is protecting the confidentiality of study participants and minimizing the risk of unauthorized data access and disclosure. To address the concern that research data will inadvertently find their way into routine medical records, it is important to fully disclose to potential participants the system that is in place to protect the confidentiality of data. If potential study participants are fully informed of the protections in place, they can make an informed decision to be part of the study or not.

When data can actually be anonymized or destroyed after use, there should be an adequate plan in place so that it can be done as early as possible in the study to minimize the possibility of breaks in confidentiality. If de-identified data are to be used, the process of de-identification should be done promptly for the same reason. If data are stored as either identifiable or de-identified and coded, procedures should be in place for periodic review of the continuing need to retain the data or participant identifiers.

DISCLOSURE OF IDENTIFIABLE DATA TO OTHER QUALIFIED INVESTIGATORS

The sharing of data is an important component of research involving potentially identifiable information, and great benefits accrue from it. Sharing among qualified researchers should be encouraged but only under authorized circumstances and under formalized processes that safeguard the confidentiality of identifiable information. Responsibility for maintaining the confidentiality of data in possession of a recipient rests squarely on the recipient of the information, not the provider of it, and this responsibility should be formalized through contractual agreements. Requests for individual patient research data for other purposes - for example, by insurance companies, lawyers, employers, the media, or other interested parties - should not be granted.

Authorization of data release to other investigators is ordinarily the responsibility of the IRB. There are at least three ways in which the IRB can exert its oversight over data release:

- Sharing data with other investigators for IRB-approved research is covered in the original informed consent under which the data were collected.

- A new informed consent for disclosure of the data is obtained from the research participant. In this case both the original investigator and the one requesting the data should notify participants that they will be contacted to have the new study explained and to seek informed consent.

- If there is minimal risk to the individual participant, disclosure of data may be approved by an IRB without requiring the secondary investigator to re-contact the participants.

Provision of identifiable specimens or derived material such as nucleic acid preparations by one researcher to another requires review, just as the original use of the specimens does. This does not mean that investigators who are collaborating on the original research require additional review. In contrast, requests for biological samples or data for new uses should require new IRB review. This might proceed rapidly following preparation of a concept that includes a description of the hypothesis and methodology to be used by the new investigators.

Researchers and research organizations should clearly document their procedures for releasing data to other qualified investigators and for accepting it from the original investigators**.** An adequate set of procedures might include the following elements:

- Only those identifiable data should be released that are necessary to achieve the goals of the proposed study. This requires a clear statement from the requesting investigator, who should justify the need for information. The specification of data elements being requested may be facilitated by use of a requestor form provided by the holder of the data. The absence of clarity in the data request may result in inclusion of ineligible individuals, the unnecessary disclosure of identifiable data, and the inclusion of unnecessary data items.

- A protocol prepared by the investigator must be submitted. This protocol will describe the nature of the study and hypotheses to be tested and the methods to be employed, with specific emphasis on the use of the identifiable data and human subject contact, if required. A grant proposal may often fulfill this requirement.

- A review and evaluation of the protocol is necessary to protect against the use of identifiable research data when non-identifiable data would suffice; to establish that the requestor is qualified to perform the proposed research and will employ reliable methodologies; to establish that the requestor is capable of protecting confidentiality; and to prevent uses of the data for purposes other than those for which they were collected without any necessary additional review (see below).

- The investigator and the party releasing the identifiable data should sign a binding research agreement. The agreement should specify the terms under which the data may be used and how confidentiality must be maintained, including sanctions or penalties for breach of confidentiality. It should prohibit re-release of identifiable data to third parties, and should delineate any other obligations of the requesting investigator. In addition to

the agreement, the party releasing the data may provide written documentation of recommended policies and practices for using the data, including relevant legal issues. Such material may help prevent inadvertent breaches.

There should be evidence of external review and approval of the data release. Ideally, an approved IRB or the equivalent should perform the review.

The researcher receiving the data or specimens should be able to provide documentation that all members of the receiving research group have been trained in confidentiality practices.

OBTAINING CONSENT FOR THE USE OF IDENTIFIABLE INFORMATION

In nearly all areas of medical research, the use of identifiable information requires informed consent. Implementation of federal regulations, a straightforward process for many years, is now beset by ambiguity in a number of scientific contexts, particularly those involving human tissue and genetic analyses. As scientific opportunity has become richer and our analytical tools more powerful, the increasing complexity of study design and the possibility of secondary uses of data and specimens not anticipated at the time they were collected have raised vexing questions about the nature of a proper informed consent.

For studies that entail minimal risk to participants, the informed consent document should seek as broad an authorization from participants as possible, provided that broad consent is consistent with imparting the necessary information effectively to participants. This is also the most convenient approach, since it tends to obviate the need to go back to study participants for more specific or follow-up consents at a later date. For example, it may be very problematic logistically to obtain consent for additional follow-up information for long-term epidemiological studies or clinical trials or for additional information to annotate archived samples. General consent for future minimal-risk epidemiologic research can often be obtained by providing an option in the consent stating explicitly that "you do not need to contact me." This allows tissue or other health information to be used in future minimal-risk research without re-contacting the subject. The alternative option that should also be provided is "I would like to be contacted in the future."

On the other hand, for many complex studies, the informed consent is better structured and presented to participants in a modular fashion. Particularly in studies wherein participants are offered multiple requests or procedures, a staged consent process, with the opportunity to consent to each request separately, may more effectively inform and educate participants than a single, complex, multi-purpose form which may be confusing. Another advantage is that the study is less likely to be compromised; a potential participant might, for example, be agreeable to a questionnaire but not the drawing of blood. The potential for re-contact to secure permission for future uses of additional information or specimens should be specifically stated. A variation on this theme is the use of a tiered consent form that offers participants the choice of whether their data or tissue will be used solely for the original research intent or for purposes in the future that

are currently unanticipated. Any of these options enables greater control by study participants over the future use of their biologic sample and research information. Such an approach honors the autonomy of participants by making their preferences specific.

It is highly desirable to avoid the need to obtain consent repeatedly from participants simply because of the passage of time. The need for frequent and repeated consent to re-authorize unchanged study objectives is demeaning to study participants, does not enhance participant autonomy, and is a significant barrier to the conduct of research. Participants should, of course, always have the right to withdraw from further participation in research at any time. Consent forms should clearly state that the participant can withdraw consent at any time, rather than specifying an expiration date for the consent. Expiration dates patronize the research participant by implicitly questioning his/her decision-making capacity.

Whatever their form and structure, consent forms should inform participants that participation in a research study means that the research team will have access to their records. The team may, for example, include physicians or other investigators, research nurses, data managers, and pharmacists. Once confidentiality training is the norm, the consent should note that these individuals have had confidentiality training and are bound by the confidentiality policies of the institution. Patients should also be notified about the extent to which their records will be accessible to officials from the FDA, NCI, drug or device manufacturers that may be sponsoring the study, or other groups. Study participants should be advised that all efforts will be made to keep their research data confidential, but that their data remain subject to the subpoena power of the courts, in the event of pertinent legal action.

A rather subtle aspect of the relationship between informed consent and participant privacy relates to the possibility of identifying study participants inadvertently from published analyses of study data. This is ordinarily not an issue because of the impossibility of identifying individuals from summary data. If, however, analysis of a particular study involves such small subsets or focuses on a rare enough condition that patients are highly likely to be identifiable, then analysis of the data should require IRB approval. Similarly, publication of results involving an individual or small subset of patients with unique distinguishing features such as pedigrees, which would make them highly likely to be identifiable, should require IRB approval prior to publication.

For tissue specimens, informed consent authorizing research use should be on record for all specimens and accompanying clinical data. This is not usually a problem when the specimen has been procured specifically for a particular research study or for entry into a research archive, but it can be difficult in the routine setting of clinical pathology, where administrative arrangements for keeping research-related records often do not now exist. It would be most efficient to combine consent for research use of tissue and associated data with the general surgical consent but to include a separate signature line permitting research use. Whenever it is anticipated that tissue will be sampled or removed for diagnosis or treatment**,** consent should, if possible, be obtained at the time of admission for in-patient and out-patient procedures, and this consent should be recorded and tracked. The actual process for obtaining consent should be left up to the physician or the institution. Certain patients may be too distracted or upset to carefully review and consider the content of these documents at the time they are presented. Some practitioners

give a copy of the consent document, or a transcript of the procedures-and-risks discussion, to take home to review at leisure prior to the actual signing of the consent.

When archiving new tissues with their clinical annotations, the informed consent for using them in unspecified research involving minimal risk should be prospective and as broad as possible, so that it will not be necessary to go back to the patient for more specific consent at some later date. The same is true for the use of medical records or other information repositories for future minimum-risk research. The National Bioethics Advisory Commission (NBAC) report permits this approach (Chapter 5, Rec. 9) but some local IRBs have required re-consent. The National Action Plan for Breast Cancer has recently developed an attractive model consent that authorizes unspecified future research. IRBs should be strongly encouraged not to require re-consent to authorize minimal-risk research.

When patients refuse or withdraw consent for the research use of their specimens, these should be annotated accordingly and not used for research. It should be possible for patients to give or refuse consent for the use of clinical data separately from the use of tissue. The default position should be “no” – that is, the sample and data cannot be used for research unless there is a positive record of consent.

Archived specimens collected prior to the current standards for informed consent should be available for research purposes, as long as appropriate steps are taken to maintain confidentiality of identifiable information. The decision regarding whether the specimens may be used should be in accord with the Common Rule definition of risk, allowing waiver of the need for consent in research that is determined by an IRB to be of minimal risk.

Consent for the research use of specimens must be maintained in the pathology department or repository records, but no notation that a specimen has actually been used should appear in the patient’s medical record.

In cancer surveillance, most research uses of data that involve contact with the participants require some form of documented informed consent. Many surveillance systems are based on mandatory collection of identifiable data. Typically, written notification that their data will be included in a registry is provided to patients receiving diagnostic procedures and medical treatment; informed consent is not obtained. Laws and regulations may further permit certain uses of these data without consent. Use and release of these data are governed by established guidelines and authorized by designated officials. Agencies responsible for cancer surveillance systems should take an active role in informing the public about the use of this information for improved public health and to advance research in cancer control. For uses of registry data that fall outside those uses authorized by law or regulation, informed consent should be obtained. Informed consent forms can either be obtained for specific uses or can be obtained to allow general access or use of identifiable data; it is probably best to obtain a new informed consent every time new data are needed from the participant for a specific study.

Historically, the analysis of health-services delivery has been an important part of care monitoring by health-care organizations: reviewing records for cost containment, quality assurance, possible negligence, and other internal purposes. When, however, such analyses are,

in fact, research, involve human subjects, are performed by an investigator, and are presumably intended for publication, the research should be reviewed by an IRB, and the individually identifiable information used in the research should be treated with the same confidentiality protections as in any other kind of research.

In family studies, there are special concerns about the future use of materials, information (including personal medical history, validated diagnoses and family history), clinical findings, and genetic test results. These can have an important impact on the health care and screening practices of other family members. Sometimes family members might benefit from information obtained from a close relative, now deceased. Consent for release of such research information can be problematic. A carefully crafted consent process which provides for difficult scenarios is important but may not anticipate all problems. There are currently no clear rules or guidelines for when previously signed consents for use of specimens or data can be considered valid for deceased or incompetent individuals, or for those whose current whereabouts are unknown. Guidelines for investigators and IRBs, along with sample genetic consent forms, are badly needed.

Making the best use of information acquired in a family study can be greatly simplified if the consent form prospectively authorizes release of information to specific individuals under particular circumstances. It is also prudent to stipulate in the consent form the specific purposes for which the donated biologic specimen can be used. A fundamental premise is that the test results and the genetic material should be available to benefit the subject's family members, and to advance the research within whatever constraints are imposed by individual family members on the use of personal data.

INSTITUTIONAL POLICIES FOR THE ACQUISITION AND USE OF DATA OR SPECIMENS

Whether data and specimens are collected specifically for research purposes or whether they are selected for use in research after having been collected in the setting of medical care, their use in research should be defined formally by the research study in question. Data should be entered into a research database and used only as part of an approved scientific protocol that has been reviewed for scientific merit, design, methodology, analysis plans, and adequacy of participant protections.

Data can clearly be used for the purpose for which informed consent was obtained. Secondary exploratory analyses of data resulting from the study that follow sound scientific principles should be encouraged; there is, in fact an affirmative responsibility to maximize use of the data. Analyses related to the original research intent should not require IRB re-review. Also, data sets that are anonymized can be used for purposes other than the reason they were collected, if no additional data are being collected from participants and the original research was approved by an IRB. For example, the NHANES data set serves many purposes not anticipated by the original collection of the data. The publication of aggregate data from such analyses involves no confidentiality issues and does not require IRB review. This practice does not cover situations where additional data needs to be collected. In that case, additional informed consent and IRB

approval is required.

A research question substantially different from the original research intent should require new IRB approval. For example, if data from a Phase III clinical trial comparing two treatments in breast cancer is analyzed several years later to explore the influence of race on response to treatment and outcome, this is consistent with the original research intent: improving treatment for breast cancer patients. However, using the same data set to examine the correlation between certain genetic characteristics and behavior seems substantially outside the original research intent and should require IRB approval.

For specimen resources data acquisition is determined by the anticipated uses of the repository. Acquisition of data for research on specimens that are not in repositories designated for research should proceed according to the Common Rule requirements.

For many surveillance systems, state and federal legislation and regulation form the basis for the collection and reporting of health data on individuals. Policies regarding who and what determine data acquisition and use should be clearly specified by surveillance programs. In the absence of legal specifications, such policies must be developed, preferably with the assistance of external advice. To maintain such policies they must also be reviewed regularly and revised according to need.

ACCESS BY INDIVIDUALS TO THEIR OWN RESEARCH DATA

Workshop attendees expressed a broad range of opinions on this subject. Patient representatives advocated that participants have access to their own data under virtually all circumstances. The great majority of researchers present thought that data should be available to participants under certain circumstances but did not favor unconditional policies of uniform availability or unavailability. Research data may be difficult to interpret or frankly uninterpretable and, by definition, have no established links to health status or medical decision-making. There was consensus, however, that any such release should be accompanied by clear language stating that the data have not been validated for medical decision-making. Also, the released data should be accompanied by names and contact information of available specialists qualified to interpret the research and answer questions about its implications.

Whatever their views on disclosure or non-disclosure for particular studies, all Workshop participants agreed that the disclosure plans of studies should be presented in the project plan that is submitted to the IRB for review, should be part of the consent process, and should be included in the educational materials relating to a study. For example, the informed consent for a research study might advise patients in the following way: “Over the course of this trial, routine health care information, such as mammogram results and standard blood tests, will be obtained and kept in your medical record. The results of these tests will be available to you from your health-care provider. In addition, trial investigators may collect certain data that are not routine health-care information, including assay results for [name of tests]; these are new cancer markers of uncertain significance. The results of these tests will not be released to you.” Special

considerations for children should be addressed explicitly with language such as: “parent or guardian has a right to look at the child’s data.”

Most research data derived from tissue specimens has not been validated for clinical decision-making and should not be disclosed to the participant under most circumstances. Such data requires validation by other studies and laboratory tests done in research settings should not be used to guide clinical decisions. In contrast, aggregate data should be made available to research participants when appropriate or if requested. In those rare cases in which providing research information would be beneficial to the patient, an IRB should review and approve the plan for reporting that information.

In clinical trials decisions about participant access to their own data should be made on a trial-by-trial basis. The informed consent process should tell potential participants about the degree to which their own data will be made available to them. When clinical trials data are provided to the patient, they should be provided by the treating physician, who will be able to answer any questions resulting from availability of the data, and not by the coordinating center. The reason is that standardized definitions of endpoints may be treated differently at the coordinating center from what was documented locally as part of the patient’s health care (e.g. definition of what is and is not a recurrence).

In considering access to data directly from registries, some Workshop participants feared that data release may violate principles under which IRB approval was granted. For example, a patient may be harmed if medical information is released to a patient from whom it had been withheld by the physician or family members. This is a particular concern for certain ethnic and cultural populations, in which full disclosure of medical details or prognosis is not normative..

In genetics research the plans for potential disclosure should be considered within the context of the type of study and the types of genes proposed for evaluation. Plans for disclosure may be very different for a family study where major genes for disease risk are being sought, compared with a large population study evaluating metabolism genes having no clear relationship to disease risk. In either circumstance, disclosure plans should describe the participant notification process and the criteria for notification. These criteria should include establishment of “clinical relevance.” In the event of unanticipated findings of clinical importance, the notification procedure for study participants should be reviewed by the IRB prior to any contact with the participants.

With specific reference to genetics research:

- If in the course of a family research study, a mutation in a gene relevant to the health of the study participants is discovered, the participants should be re-contacted. Notification of these results is only appropriate when the clinical relevance of any finding has been established and meaningful education and counseling is possible. The American Society of Clinical Oncology has published practical guidelines for this process (J Clin Oncol 1996;14:1730-6). Participants should be informed that a mutation related to the disease being studied has been identified in the family. They should be offered the opportunity to

learn their personal mutation status if they wish. Some may decline. If they are interested in clinical mutation testing, they should receive appropriate education and counseling before a new biologic sample for DNA is obtained. Consent for genetic testing must be done on an individual basis. Consent cannot be done in a group setting without risking coercion of some individuals. Notification about the results of mutation testing should not be based on the testing of an archived sample. A new sample should be obtained and should be tested in a laboratory that is certified by the standards of the Clinical Laboratory Improvement Act (CLIA). The results should be given to the participant in private with appropriate counseling. Test results should be delivered without the involvement of interested family members if possible. Risk notification, including education, counseling, and CLIA testing may require supplemental funding from the funding agency.

- In large epidemiologic studies, it is often not possible or feasible to re-contact study subjects with individual genetic testing results, especially if the testing was done on an archived specimen. In this situation, we recommend that a newsletter informing individuals of the aggregate results be sent to study subjects so that they have the opportunity to pursue the findings individually if they are interested. Again, supplemental funding may be necessary to cover these costs.

- Individuals participating in federal research are covered under the Privacy Act and therefore have, by law, access to all information collected from them. However, access to all research information is not always helpful to individuals. For example, in genetic epidemiological studies investigating gene-environment interactions, genes which have uncertain relevance to a study participant's health may be studied. Even in family studies where a gene directly related to a specific disease is being sought or investigated, the actual relevance of mutations in that gene to the study participant's health or future risk of disease may not be clear. Disclosure to a study participant that he or she carries a genetic alteration of uncertain significance may not be clinically relevant, and may actually cause harm (stress, anxiety, depression). Most of these studies use archived biologic specimens, for which there is a greater chance of sample mix-up than in fresh samples which go directly to a CLIA certified lab. The subcommittee agreed that all reasonable efforts should be made to educate research participants about the difficulty in interpreting (unvalidated) research data of uncertain significance. If the participant still requests access to genetic data of uncertain significance, their personal information should be divulged along with referral to local counseling resources to help explain the implications of the information.

- These concepts should apply to all research participants, including minors, parents of minors, and others unable to consent and their guardians.