

MINUTES

ALS RG RESEARCH GROUP MEETING

ALSRG DATABASE AND DNA BANKING

October 6&7, 2005

(prepared by Petra Kaufmann)

1. State of the ALSRG (Hiroshi Mitsumoto)

The ALS Research Group (ALSRG) meeting, with sixty-seven members in attendance representing 51 of 76 active sites, convened on October 6, 2005, in Chicago, Illinois, with an update on the state of the ALSRG. The by-laws and constitution have been developed, ALSRG officials have been elected, and administrative committees have been formed. The short term goals of the ALSRG include an ALS biomaterial and data repository. Long term goals include the facilitation of therapy trials, interactions between ALS investigators, and awareness of therapeutic opportunities. The ALSRG will also foster the recruitment and development of basic scientists, young investigators, and clinician investigators.

In the initial phase, the ALSRG will focus on a DNA Repository and Database. The NINDS will provide a supplement for funding for a repository and is welcoming application in a coordinated effort such as the ALSRG is considering. The MDA and ALSA have indicated their willingness to provide additional support for an ALS DNA Banking initiative.

2. Complex Genetics in ALS: Towards a Whole Genome Scan (Robert Brown)

The last 5 years have seen great advances in new technologies that provide greater opportunities for analyzing genetic factors in ALS. For example, in Mendelian Genetics one looks for linkage between phenotype and genetic markers. In recent years, several genes have been identified, only two of which cause "true ALS". Complex genetics research studies the overrepresentation of genetic variants in the ALS population as compared to the control population. The assumption is that something related to this marker somehow sets the stage for an individual to develop the disease or to possess the specific phenotypic characteristics of the disease. One can compare alleles, genotypes or haplotypes between the disease and control groups. The next step is to identify the biological significance of differences in genetic markers. One has to consider the distribution of the markers in samples from both control and patient populations and the overlap between them. The closer the means of the case and control populations, the greater the population sizes that will be needed to guarantee an appropriate power to detect a given difference. Therefore it is important that ALS investigators collaborate in collecting the critically important large number of samples. To detect smaller OR's (e.g. about 1.2) one may need greater than 1500 DNA samples. The acquisition of DNA (data) is the first step and the NINDS Supplements will provide a unique opportunity to catalyze the acquisition. In parallel to the ALSRG/Coriell initiative the European UK BioBank has also started a biomaterial collection. In the Whole Genome Analysis Project (WGA) there are multiple potential participants for an ALS project, in the US, Canada, the Netherlands and other countries. This will provide a unique opportunity to study the etiology and pathogenesis of ALS, and there is precedent from the study of other diseases that this kind of initiative can be successful.

3. ALS DNA Banking – Program Description (Edward Kasarskis)

Sites participating in ALS DNA Banking will submit a blood sample and a short data form to the Coriell repository. Coriell will then create cell lines and bank DNA as well as enter data from

forms. The biomaterial and data will then be available to other investigators. The supplement mechanism through NINDS sets the framework, which is as follows: The sites of ongoing NINDS funded research studies will serve as lead sites that will help coordinate the collection effort through their consortium sites. The projected timeline for the next two years is to collect 4,000 samples (2,000 ALS and 2,000 controls). When 25 pairs of samples have been collected per consortium the NINDS will reimburse the lead site. Given the total amount of funding, this would result in 667 samples per NINDS lead site and group over 2 years, and 333 per NINDS lead site and group each year. Thus \$ 600,000 in funding would result in a total of 4,000 samples: 2,000 controls and 2,000 ALS samples.

A single master protocol will serve as the organizational unit. The individual sites will each submit the master protocol to their IRB and will enter into a subcontract with the lead site. The sites will collect samples and ship them directly to Coriell. They will also submit a one-page clinical data element (CDE) form to Coriell and notify the lead site of the transaction. The lead sites will participate in sample collection, coordinate the subcontracts, track progress, document blood and CDE submission, submit Request For Payment to the NINDS and reimburse the individual sites for work.

Individual sites will receive \$150 per sample and completed CDE form. Indirect costs are not applicable to this supplement as the “per patient reimbursement” will be for phlebotomy, a patient care item, rather than for research or administrative functions. The Lead Sites will each apply to MDA or ALSA for a 50% of a FTE coordinator.

4.) Minimal Data Set “Short Form” (Alex Sherman and Robert Miller)

During a prior meeting, the task force discussed what data elements would be necessary to constitute a minimum dataset that would complement efforts to elucidate the complex genetics of ALS. The first step is to collect information summarizing demographic and basic clinical data. This minimal dataset, similar to the datasets that have already been developed for use in other diseases such as Alzheimer’s or Parkinson’s diseases, would not be sufficient for epidemiologic research or similar efforts. Thus it is important to realize that this dataset is a dynamic one. The expectation is that the data collection document would be revised periodically. Information is needed so that databases can be searched for mortality, however due to privacy requirements, individual information cannot be included.. Information collected will include: clinical site, subject ID number (specific to this study and site), first three digits of zip code, racial category, month and year of birth, date of onset and site of first symptom, ALSFRS_r, FVC, date of diagnosis, riluzole, PEG, NIPPV, asst vent >23 hours and date started, and tracheostomy. The form also asks for signs supporting diagnosis. There will be options to determine if upper and lower motor neuron signs are present, absent, indeterminate, or not tested in a given region. If EMG studies are available, a copy must be attached. Information required pertains to whether or not there was evidence of acute or chronic denervation in the four regions. There is also a place to state the level of diagnostic certainty per EEC. Further, if there is a known mutation, it can be documented on the form, along with any atypical features. Lastly, the investigator, an ALS expert, verifies and signs to indicate that consent has been obtained and that the information is accurate and the ALS diagnosis correct. For this study, there is no “embargo” time, thus samples will be available immediately. The issue of fronto-temporal dysfunction was also discussed here and likely will require further definition.

5. Web-based Data Collection at the Coriell Institute (Judy Keen)

The Coriell Institute is currently accepting samples for ALS and has an on-line form for ALS which will be updated at the recommendation of the ALSRG. Coriell provides blood tubes, shipment containers and covers shipping costs. The samples are sent at room temperature. When a sample arrives at Coriell, a unique number is assigned to it. In addition, at the site there is a

local ID number that the site can use when tracking a sample. Demographic and diagnostic data can be entered and updated via paper form or alternatively through a web-based system which shows data that has already been entered for a given subject. On the Coriell Institute web-based data repository, drop-down menus are available to facilitate data entry, for example, giving choices for secondary diagnoses. A data dictionary is available separately to define the criteria for the given variables and there is also the possibility of “help” screens. Longitudinal data can be entered through link of local ID. Once the ALSRG has finalized the CDE form, Coriell will update the web-based form within one week. Coriell is frequently in contact with submitting investigators to complete missing data points and confirm that the data form is finalized. The group discussed whether all information collected in this “short form” is necessary for the genotyping effort and in particular if longitudinal data are worth the effort. The investigators argued that the current “short form” is already a condensed form and that this information is very important to characterize the phenotype including disease progression. PEG, tracheostomy, etc., and the approximate times that they were implemented, in conjunction with disease onset, give estimates of disease progression, and were included in case information on mortality was not available,

6. Site Benefits and Responsibilities (Jeremy Shefner)

Site responsibilities include the following: submitting IRB Application and Informed Consent to local IRB, obtaining blood samples from ALS patients and controls, providing complete information, and sending materials to the central facility for banking. The benefits for the sites are to be a part of the first large scale DNA banking project in ALS and to contribute to the vitality and future of the ALSRG. Finally, each site will receive \$150 per sample/data form. No indirect cost will be taken off this amount. This is similar to the model of ALSA grants where indirect cost is not provided. Academic benefits include individual investigation in that investigators may purchase DNA samples at greatly reduced costs (\$10 per sample). For group investigation as a consortium, the ALSRG will have access to the total DNA repository at a reduced cost. There will also be recognition of the ALSRG and individual sites for all studies performed by the group. Finally, there will be recognition for submission on the Coriell website.

7. Logistical Issues (Petra Kaufmann)

An IRB protocol and consent forms have been prepared for use of the group. They highlight the significance of this research, the study organization, the flow and availability of biomaterial and information, confidentiality, property rights, risks and benefits, and eligibility. ALSRG participants are invited to provide information on what may be required from the IRB at their sites. A revised draft will be provided to participants for review.

Discussion: The group clarified a number of issues including that control subjects must be healthy and cannot have first degree relatives with a known neurological disease. This raises several issues. First, control subjects will likely only be seen one time by the investigator. If a neurological disease develops in them or their relatives following the blood drawing, the investigator will likely not find out. Secondly, if ALS patients are recruited regardless of family history of neurological disease, but patients are recruited with family history of neurological disease, this sampling strategy is problematic from an epidemiological point of view.

1. State of the ALSRG Website (Eric Sorenson)

Dr. Sorenson showed the website he created and solicited the group’s feedback. He has created a logo showing a motor neuron and the group’s name. Steering committee members are listed on the website

without their email addresses. There are links to other relevant groups, as well as relevant documents and forms for the group, including the “short form”, the group bylaws, and the ALS member list. The website is designed to be informative for ALSRG members, rather than a resource for the general public.

2. ALSRG Data and Biomaterial Repository (Robert Miller)

The “short form” was modified with the input of the group and will continue to be modified. One issue that was raised is the selection of controls. The criteria were developed by the NINDS in order to make the control group useful for groups studying different diseases within the Coriell Repository. In order to have more comparable information on patients and controls, information on other neurological diseases will be added to the family history of patients. Also, the definition of fronto-temporal dementia (FTD) will be clarified and provided. Both prior and future longitudinal measures, although not obligatory, will enhance the quality and value of the data. These additional data can be entered at any time on the website and will not be required in order to submit a sample to Coriell. The hope is that given these comments, the ALSRG membership would entrust the further modifications to the Oversight Committee so that the entire package could be submitted to the NINDS shortly. The group discussed how FTD should be ascertained on the CDE given that the clinician’s impression alone is unreliable. Others argued that collecting information on the “clinical impression” on cognitive function may be the only feasible method because few investigators are familiar with the FTD criteria. Regarding a participant’s ability to give informed consent, the group discussed that a legal guardian would have to give consent for those severely impaired.

The group reviewed access to the samples: If the group submits 2,000 samples and then endorses a project that uses 2,000 samples they will be able to purchase the samples at the reduced rate of \$10.00 per 10 microgram. If, however, the group requests more than 2,000 samples, they would have to pay the higher rate of \$100.00. It is important to consider that the importance of having access to the samples outweighs the cost of obtaining them and any future research grant proposing the study of ALS or control DNA samples will have to budget for the sample acquisition. Thus, the first group projected would have to be carefully considered by the group as it would be the only project benefiting from the low cost access to the number of samples that the group has submitted. It was pointed out that the DNA can be amplified for future genotyping, so that once the group has purchased samples they can be maintained for future use, however, it is unknown if this is possible for certain genome wide searches using chips because amplification might cause mutations or because of additional methodological concerns. This repository has its limitations, but is an important first step in creating a valuable resource. The group voted unanimously to proceed with collaboration with NINDS/Coriell under the suggested organizational structure through the six teams.

3. Tissue Banking (Jean Paul Vonsattel)

The aims, procedures, and experience with brain banking were reviewed and it was pointed out that there has to be a 24hour/7day call system in order to have optimal research samples. Aims include collecting brains as soon as possible after death for processing by experienced processors, and the latest techniques include genomic proteomic technologies. The availability of these samples is essential for investigating the mechanisms causing premature, neuronal death in neurodegenerative disorders. Brain banking was initiated in the modern sense in the 1970’s when the notion was accepted that only one half would be studied pathologically, and the other half would be prepared for research studies, thus assuming that disease is similar in both halves of the brain. It is crucial that the dissection be done by experienced pathologists and that the instructions on where to cut are simple and followed precisely. Freezing techniques are also very important so that morphological features are preserved. A preferable method similar to those used in fertility clinics is used by taking liquid nitrogen vapor at -180 to -160 degrees Celsius, as immersing the specimen into liquid nitrogen where breaks may occur. A consensus was reached on which blocks should be obtained on dissection and

how they can be tracked. Since January 2002, at the bank at Columbia University has processed 46,871 CNS samples and disbursed 3,942 specimens. For the purpose of expedited and appropriate disbursement of specimen they have developed an electronic software system using bar coded labels. Sixty tags are prepared in advance, and specimens are submitted using those tags. Matching samples are re-selected. This system gives the coordinates of the samples' places in the freezer and tracks the disbursement of samples and vacant freezer space upon disbursement so that new samples can be stored in those spaces. The recipient investigators are registered and the NYBB software keeps a current inventory of stored samples. The freezers are safeguarded by a back-up system which includes one empty and running freezer that is constantly available.

Once a sample is categorized using the bar code system, any sample can be retrieved from the brain bank in less than 2 minutes. The tissue will be disbursed to the investigator who asks first and who has IRB approval at his or her own institution, after the physician who treated that patient and who had arranged for tissue donation gives approval. The brain banking system is global, but has special protocols for example for Alzheimer's Disease or Amyotrophic Lateral Sclerosis. The brain banking is funded by the Taub Institute, Nancy Wechsler, and Parkinson's Disease group, among others. Additional funding would be needed, if the brain bank were to be used more frequently for ALS. The IRB issues are less stringent than for living patients, essentially requiring that the patient or patient's family have given permission (signed) for autopsy material to be used for research and secondly that the tissue is de-identified. The brain bank contains control brains, donated either from autopsies performed at Columbia University Medical Center for non-neurological reasons, from the Northern Manhattan Aging project which follows elderly non-demented subjects, and from other centers. The bank works with Sterling courier, a company that sends a car to the institution where autopsy is performed and brings specimen to airport. The specimen is then brought immediately from the airport to Columbia University Medical Center. This infrastructure was found to work very well. The time from autopsy to arrival at the brain bank seems less important than the condition of the patient at autopsy.

The cost of an autopsy is about \$ 4,000 and is sometimes paid for by the Department of Pathology or the ADRC (Alzheimer Disease Research Centers). Dr. Brown thinks that it is desirable to use existing infrastructure in order to organize tissue banking for ALS. Transportation within the US is \$300-400, but allows for optimal samples. In addition there is a cost for transporting a body from a home or local facility to a medical center where tissue can be retrieved, which in New York City can range from \$600 to \$1,000. The importance that the interval from postmortem to autopsy is as short as possible and that cryosections are obtained immediately, especially for RNA studies, was stressed. It was also pointed out that the current multiple efforts to collect ALS tissue are localized and segmented and could be coordinated better.

4.) Center based versus population based epidemiology in ALS (Lorene Nelson)

Population based studies are expensive and require sophisticated infrastructure and thus may not always be the best approach. It is important to remember, however, when considering different study designs that there are some important differences between patients seen and those not seen at specialty clinics. The external validity (generalizability) is less important than the internal validity. Two population based studies that speak to issues of referral bias were mentioned, one from the United States (Lee et al, J Neurol Sci 1995) and one from Ireland (Traynor et al, Neurology 2002). In both studies the average age of patients seen at specialty centers was lower than for patients in the community. Among those seen at specialty centers, there were more male patients, fewer patients with bulbar onset. Also, riluzole use was higher, NIPPV use was higher, and the survival was longer in those seeking care at a specialty center compared to patients in the general population. Patients seen at specialty centers are also more likely to have familial ALS. However, once the differences between specialty clinic populations and general populations are understood, important observations can still be made using a center-based study design.

It is hoped that the NINDS will reconsider the exclusion criteria for control subjects as this may cause problems in the restrictive criteria for validity of the study design. Under-representing subjects with a familial history of other neurological conditions, under the current design, may lead to the detection of a false association between cerebral vascular disease and ALS, as, for example, those with family history of stroke are excluded from the controls.

If the study goals are to identify highly penetrant disease-causing mutations associated with familial ALS, this is best done based on multiplex families. These families are typically already being referred to researchers that can do these kinds of studies as this approach still seems to be the main way to find susceptibility genes for familial ALS. A more applicable goal for the repository effort is to identify susceptibility genes for sporadic ALS by looking for gene “main effect”, gene to gene interaction, and gene to environment interaction. This can also be achieved through association studies using a case-population control design or a case-family control design, as well as a cases only design. The assumption here is that the gene is not associated with the exposure or the environment, which may be the case for genes influencing behavioral characteristics (violating the independence assumption). A third goal is to examine the association of genotype with phenotypic features, treatment response and prognosis. These are cross-sectional and prospective studies that use specialty center patients or those in trials that need to complete follow-up and can be done in referral centers. Social security numbers should be collected on all patients at all sites as this will allow for the future possibility of contacting the National Death Index to obtain the date of death for a given patient.

Association studies basically compare the proportion of cases and controls that have the genotype, haplotype, or disease-associated SNP being studied. Even though the specialty center cases may be different from those in the general population, this can be ascertained for the cases and still be valid. The control selection, however, is perhaps more critical here. A good approach is to think of a “gold standard design” that answers a given research question and then to come up with a feasible design that is as close to the “gold standard” design as possible. The gold standard design, however, can only be done exceptionally, for example in Scandinavian countries, by HMO’s like Kaiser, or in a census region that is based on a well defined, enumerated population.

In reality, many studies involve samples from undefined populations and possible sources for controls would include spouses, other relatives, friends or neighbors, or neurological or general practice controls with other diseases, such as those with headaches or back pain. Control subjects drawn from general neurology or other hospital settings can falsely overrepresent characteristics that are associated with poor health (e.g., cigarette smoking). Also, when using controls with benign neurological conditions, they may come from geographic areas that are different from those of the patients who have serious conditions because the latter may be more willing to travel long distances to the specialty center. The problem with using friends or spouses as controls is that they are overmatched with respect to lifestyle characteristics. In short, when investigating environmental risk factors, these are not the ideal controls; however they would be appropriate when looking at genetic risk factors.

Principles for selecting controls include choosing those who represent the prevalence of the genotype (or risk factor) in the population from which the cases arose. Also, one aims to choose those who are likely to have navigated the same set of “selection forces” as the ALS cases, i.e. those who likely would have come to the specialty center if they had developed ALS. Lastly, the same inclusion and exclusion criteria have to be applied for cases and controls and controls should be matched for gender and age.

The group potentially has the opportunity to conduct population based studies on a subset of referral centers within the ALSRG. Methods to identify cases for population based studies include disease registries, populations that provide sampling frame (HMOs), or those that combine the methods described below. Convenience samples use general neurologist, physician, or university referrals, specialty centers, patients and service organization rolls (MSA, ALSA), patients in nursing homes or long-term care facilities, and finally, death certificates.

A suggestion was made to collect information on the source and nature of control subjects at each site, placing higher priority on selecting nonbiological relatives of neurological disease cases (not limited to ALS but possibly including relatives of Parkinson's Disease patients). The suggestion was also made to apply the same inclusion and exclusion criteria for cases and controls (e.g. family history of neurologic and psychiatric disorders). Also, follow-up plans are needed for tracking patients if the goal is to examine genetic factors associated with survival. In order to achieve a balanced case-control distribution with respect to gender and age in the analysis, the number in different age strata must be balanced. When adjusting for age, one estimates an odds ratios in an age strata and if there is a large imbalance in a given strata, then the confidence intervals become very wide.

It was reported that the ALSA has funded an ALS Consortium for the Epidemiologic Study for ALS, which will serve as a methodological resource as opposed to a database. The objectives of this consortium include: 1) to form an epidemiology consortium as a methodological resource for investigators to conduct research on the environmental and genetic factors associated with sporadic ALS, 2) to lay the groundwork for future research collaborations so that risk factor data can be pooled across studies, and 3) to develop a consortium web-site where investigators can obtain information on methodological approaches, risk factor modules, data dictionaries, and data quality control protocols. Several Parkinson's disease study groups have recently tried to combine their data from different studies to look at smoking as a risk factor and this took three years as the questionnaires that had been used could not be easily combined. The consortium is developing a website which will summarize the epidemiological literature on ALS, present studies that have been done, and make instruments available to investigators which will include standardized data collection instruments and data dictionaries, etc. The specific aims of the consortium include the following: 1) to develop a consensus on what standard data elements should be collected in epidemiologic studies, 2) to develop standardized data collection instruments for assessing risk factors for ALS, so that data items can be added easily as research questions evolve, 3) to develop computerized data collection forms that can be used with risk factor modules, and 4) to track data elements across studies to enable alter collaborations. The ACES initiative is in contact with the EURALS group to see if collaboration is possible. The EURALS group uses different case definitions than the EEC which are not always applicable to epidemiological studies.

5.) National ALS Registry Initiative (Lucie Brujin)

Due to the success of the VA registries, many patients want to be a part of a national registry and as a result advocacy groups are promoting a congressional bill that would establish such a registry. The concept is that an organization like the CDC would house a national registry and that data and any research coming from it has to be based on input from the ALS community. This would only be a minimal dataset and would have some overlap with the minimal datasets developed by the ALSRG. The ALSA is supportive of the ALSRG efforts.

6.) ALS Care Database (Robert Miller)

The ALS Care Program was started in 1995 and since that time forms have been modified and patients have been enrolled. The database was designed to look at and improve outcomes and to educate patients, the public, and health care providers on ALS care. It allows assess to practice parameters, at least at large centers where that data is being collected, and is an observational cohort study. The data collection instruments are standardized and there is broad participation across the country to collect data confidentially at approximately six-month visit intervals. The Medical Advisory Board has oversight on all aspects and includes scientists, neurologists, nurses, and patient representatives. The sponsor has no influence on the content and operations of the database. The study coordination center at the University of Massachusetts has extensive experience with large outcome research databases and has a confidentiality certificate. The data collection instruments contain physician-reported, patient-reported and caregiver-reported data. The physician reported data includes a minimal clinical

dataset. The patient self-reported data takes at least 20 to 30 minutes to complete, but most patients are motivated to fill out the form. The current enrollment includes 325 clinics with 110 clinics submitting data, with the majority of data coming from 15-20 centers. The challenge is that many patients are enrolled on a one-time basis and follow-up data is incomplete. Quarterly reports provide feedback to the participating centers and are confidential. The database is observational and there is no control group. The National Alzheimer's Coordinating Center Database serves as an example of a successful database using website capabilities. The database did not start with new cross-sectional data, but contains data that had already been collected. It started with an initial dataset (61 elements) and then 9 datasets or modules (900 elements) were added on thereby enlarging the number of variables. For all studies that were included, data were posted through this mechanism. The data areas include the minimum dataset ("short form"), neuropathology data, centers' data, and collaborative data. It also contains a data element dictionary as well as metadata (data about the database). Recently, this database has allowed for a study of APOE and AD from patients included in the database and the data became available to investigators with permission from the larger group. There are 4 security levels depending on one's position and role in the organization. They also have analysis files versus customized reports, which creates tables on the website that can be downloaded using excel. They have 66,000 patient ID's including neuropathological data on many. The ALS Care Database leadership has experience and is eager to possibly serve the ALSRG when establishing any future database. The ALS Care database also has received unrestricted funding from a pharmaceutical company. The leaders of the ALS Care program are open to new governance, possibly an ALSRG committee to replace the current Medical Advisory Board of the ALSCARE Database Project. The ALS care program is open to everyone, but currently only appeals to a relatively small number of sites. The group pointed out that the level of participation is related to the accessibility of the data for research by others.

7. ALSRG Databases and DNA Banking. Integration of distributed semantically heterogeneous data sources (Alex Sherman)

Potential projects for the ALSRG include DNA banking, DNA banking extension projects, inventory of ALS Data Sources, ALS disease ontology creation, and integration of heterogeneous data sources. The ALSRG can benefit from the structure of Coriell, but should not rely solely on Coriell and should have a data repository that is independent from other efforts. A separate ALSRG database must consider how it can be synchronized with the dataset that is submitted to Coriell. Servers, connections, permission, validations, data elements, and dictionaries have to be developed and defined so that they can continuously and periodically synchronize different data sets. For the DNA banking project, mechanisms are also needed to track samples and develop payment systems. There are important identifying data elements, such as social security numbers, that would allow for crucial follow-up information; however these elements cannot be collected in a central database, and must be maintained at the local site. Once the local site has made the required query to a national database, for example, the National Death Index, the result, in this case, survival time, can then be reported to a central database. Certain extension projects to the ALS DNA Banking effort may require additional data collection and cross-referencing between DNA banking and other data sources is important. Thus, when building a database future needs must be anticipated. An inventory of ALS Data Source is also needed and may include registries, clinical trials databases, and other clinical data sources. Issues here include ownership, access, credit, confidentiality and regulatory compliance. There are new trends in data management that may be of interest to the ALSRG: Semantic web is a project and was initiated by the creator of the world wide web. The underlying idea is that information is available not only through URLs but also that objects are assigned web identity. Ontologies define both relationships and elements, whereas data dictionaries define only data elements. There are two important technologies, extensible markup language (XML) and resource

description framework (RDF), the challenge being to provide a language that expresses both data and rules for reasoning about the data. XML allows for the addition of arbitrary structure but says nothing about its meaning. Meaning is expressed by RDF (encoded in sets of triplets- subject/verb/object) – which states that certain objects have properties with certain values. Subject and object are each identified by a Universal Resource Identifier (URI).

It would be desirable to simply create a new dataset into which existing databases could be “dumped”, however this is not possible as there are many technical obstacles and data elements that are not easily comparable and transferable between databases. It is possible for the ALSRG to contract with a vendor or to create tools that would help the owner of a database to submit a dataset for common usage that would allow for the accurate transfer of data and would be fully accessible to all, not only to the contributors to database. Careful planning is necessary to maintain confidentiality and to make such a dataset compliant with all regulatory requirements.

Challenges in integrating data sources include large numbers of data sources, unavoidable differences in data definitions, and the need to retrieve and analyze data from multiple sources. The effective use of such data would require the reconciliation of semantic differences among the relevant data sources. There is also a need to develop tools that would allow for ad hoc queries and that would meet the needs of the research community.

Solutions to these problems potentially include creating data source-specific information (data ontologies and data schema) for existing data sources and defining procedures for a data source registration through definition of its schema, location, type and access procedure. Resulting ontology-extended data sources would allow for the specification of semantic correspondence between the user ontology and the data source ontologies by specifying inter-ontology mapping. The advantage of this approach is that it would not be necessary to organize a central server or back-up mechanisms, etc.; however, it would require data owners to manage access to and security of their data source. A collection of autonomous, semantically heterogeneous distributed data is available as a set of inter-related tables structured according to a pre-defined ALSRG ontology. The next steps include defining financial and human resources, priorities, projects, and vendors.

8. ALS Registries (Benjamin Brooks)

Currently, State and National Registries include three regional registries: Olmstead County (Minnesota), Washington State (three counties) and Bexar County, Texas. Other registries are being considered including the NYS registry.

The public health information system is an information network facilitating these efforts. There are several components to a registry or surveillance system and a module for a CDC surveillance program is being developed. The health alert network (HAN) is a mechanism that will work up outbreaks and attempts to have a national disease surveillance system. Recent epidemiological reports have shown that the incidence of ALS and MS has risen over time and may be more common in Northern latitudes. If an unusual increase in a disease is noted, early detection activities are initiated.

Input from the group was welcomed and investigators were invited to join the database committee if they are interested. It is anticipated that the ALS Banking project will move along rather quickly, but that any extended data collection will require further preparation. The ALSRG Database Taskforce is meant to include “new and young members.”

9. Expanded Data Collection “Long Form” (Robert Miller and ALSRG)

The ALS DNA Banking will be a good initial project and the oversight committee can consider developing a database/tracking system for that project. An incremental approach was suggested where one starts from the “short form” (common data elements, CDE) and then adds a limited number of additional elements at a specified time agreed on by all. Once the agreement is achieved, the next set of data would be considered. A modular approach was suggested whereby specific areas of interest

would be identified and smaller groups assigned to develop modules to address these issues. The modules would then be used to collect data in addition to the “short form”. Existing databases must also be included to ensure that no information is lost. Database owners must be willing to share data and to inform the database committee what kind of data they have in their respective databases. Coriell is flexible in terms of adding data elements, but can never store any identifiers. The issue of patient requests for genetic testing results was raised. At this time, the results are for research purposes only and will not be shared with patients, as is stated in the consent form. The grant application will be sent in the near future and the impression is that biosketches and letters of intent will not be needed at this time.

Action items include

- a. Dr. Miller will send final CDE. Everyone has one week to respond and Dr. Kasarskis will finalize the CDE form, through the DNA Banking Oversight Committee (Drs. Kasarskis, Brooks, Keen, Sherman, and 6 NIH PIs).
- b. Dr. Kasarskis will find out whether biosketches and letters of intent are needed through Dr. Katrina Gwinn-Hardy and get back to the group. Dr. Kasarskis will prepare grant application within 3 weeks
- c. Dr. Kaufmann will send IRB protocol and Consent form to group. Everyone has one week to respond.

The next meeting will take place in Dublin, Ireland on December 7, 2005. A similar meeting is planned during the AAN meeting in San Diego, CA. An additional meeting may be needed next fall. A majority of current attendees indicated interest in a free-standing ALSRG meeting similar to the current Chicago meeting in a poll. The topics for the future meeting will have to be determined as a result of the committee work.