VIII. DATA MANAGEMENT

A. OVERVIEW

The data management system used in the Type 1 Diabetes Genetics Consortium (T1DGC) consists of two components. The first is processes for data collection, data entry, specimen tracking, quality control and management. In addition, the Coordinating Center will regularly monitor data completeness and timeliness and provide feedback to all Regional Network Centers who will, in turn, provide feedback to data collection sites to ensure that study data are complete, accurate, and collected in a timely fashion. The Coordinating Center will also facilitate necessary communication between the clinic sites to the laboratories for rapid feedback.

The second component of data management relates to the genetic data. The Wake Forest University Health Sciences group has extensive experience as a genetic data coordinating center, serving the Collaborative Study on the Genetics of Asthma (CSGA), the SCOR for Genetics of SLE, and the IRAS Family Study. Genetic data from multiple genome scans now resides at the Coordinating Center. Integration of genetic maps based upon existing public databases is required with each additional genetic experiment. Data from CIDR, MHC Fine Mapping, Rapid Response, Genome-Wide Association, and TaqMan Laboratories will be integrated into the database. With recruitment of 2,800 new ASP families as well as trio families, cases and controls, multiple shipments will be required. Upon receipt of these data, semi-automated pedigree checking and error detection systems are invoked. Only after evaluation of markers, allele frequencies, map length and integrity are finalized will data analyses begin. Once data have been analyzed and results are considered to be "in press", the Coordinating Center will make the data publicly available by request to the T1DGC Access Committee. This process has already been implemented with the US/UK data recently published (1).

B. DATA MANAGEMENT AND QUALITY CONTROL PROCEDURES

1. Data Entry Certification

Prior to performing data entry, each data entry person is required to complete the certification process. This requires that each individual enter two sets of data forms into the system. Once these forms are entered, they will be verified for accuracy and scored. A minimum acceptable score of 99.5% must be obtained prior to entering study data. Once an appropriate score is reached, the individual is provided access to the "live" site for data entry purposes. If an acceptable score is not achieved on the certification exam, the individual may again re-attempt certification after further training.

2. Edit Checks

Computerized data validation routines are used to enhance data quality. These include, but are not limited to: initial screening of data, using logic and range checks built into data entry screens and cross-form functional and consistency checks.

The data entry screens require a minimum number of completed fields before a form can be accepted. Validation checks are applied during data entry. Insofar as possible, checks are programmed using Rules Engine Software and corrections made as the data are entered.

A more sophisticated series of checks will be made after the data have been entered. Computer edits will be performed across forms to detect and correct instances of entry and transcription errors that pass the cross-sectional (intra-form) logical and range checks of the data entry screens. Online reports and the T1DGC query system will list these errors will be available to the Regional Network Center for verification based on hard copy records of forms or clinic information. When errors are discovered in the data, corrections will be made to the hard copy of the form in accordance with the *T1DGC Manual of Operations* and the database will be corrected by the Regional Network Center.

3. Disaster Recovery

The primary data holdings will be maintained at the Coordinating Center, who will ensure that routine data backup occurs and available if, for any reason, there is need to restore data. All data, programs, code and documents associated with the T1DGC project are backed up to a digital linear tape (DLT) tape library every night. These tapes are kept indefinitely and are located in a fireproof cabinet that remains locked at all times. Periodically, copies of tapes are moved to an off-site location for storage. In the event that there is loss of any data, the information can be restored from tape in a matter of hours. The entire computer facility is provided with conditioned power, uninterruptible power supply (UPS) capability and environmental sensors with notification protocols.

4. Security

All data entry for the T1DGC project is done via the web-based data entry application. Documentation of the data entry system will be maintained at the Coordinating Center. Online reports relating to recruitment goals, data edits and quality control will be available via the web site.

Normally, data are transmitted across the Internet as plain text. It is possible, though highly unlikely, for someone to monitor this traffic, and using the proper equipment, reconstruct the individual pieces into the original data. Due to this threat, a digital server certificate from Verisign, Inc. is in place. This certificate allows communications between the web server and the client system to be encrypted. This encryption is as advanced as is now allowable by the United States government. This is the same mechanism used by the banking industry and for electronic commerce. This system provides more than adequate security against unauthorized use.

The Coordinating Center is protected by a Cisco firewall that limits the source and type of traffic coming into the institution. This product remains under constant monitoring and control.

5. Passwords

Access to the web site is controlled by user authentication. Each user will be giving a username and password and will have only have access to specific areas of the web site. The user should under no circumstances share their login information with anyone.

6. Quality Control

The Coordinating Center will verify the accuracy and the integrity of the data that are entered into the tracking system and the web-based data entry system. Examples of quality control procedures in place are described below.

a. Forms will be readable. Particular attention will be paid to spacing, print size, and print type. Response fields will be close to their corresponding questions so that there is no ambiguity as to which response should be paired with which question.

b. Duplication of requested information will be limited. Data collection tools will be evaluated to ensure that as much as is possible the same information is not collected more than once. It may be required, in some extreme cases, to collect the same information more than once, but careful attention will be paid to this issue, in order to keep this to an absolute minimum.

c. Instructions will be clear and concise. Various short instructions will be included on the forms when needed. Any questions that might arise during the data collection process will be resolved via the line-by-line (QxQs) instructions in the *Manual of Operations,* and if still unresolved, solicited to the Regional Network Center.

d. Consistency in response coding. Responses will be coded consistently within and between the various forms.

e. Forms will be designed that require a response for each question. There will be a "not applicable" response so that no fields will be left blank. Skip patterns will be used on a limited basis.

f. Response categories. The response categories for data that are not numeric will be mutually exclusive and exhaustive, unless otherwise specified.

g. Numeric data units. For numeric data, units will be specified and the proper amount of space will be provided for valid responses.

C. DATA

1. Clinic Data

Data collected at the clinics will be recorded on paper forms. These forms will be checked for consistency at the clinics prior to being shipped to the Regional Network Center for data entry. The Coordinating Center has developed a sophisticated web-based data management system that will be used by the Regional Network Center and laboratories to enter the data from the paper forms. Once the forms are ready for entry, the user will use Internet Explorer to access the web-based application at the following URL: https://www.t1dgcdataentry.org.

After reaching this site, the user will be required to login to the system. The process of user authentication will allow the system to determine who the user is and what access they have.

The system supports varying levels of security and when each user is created, they are assigned a specific level of security based on what role they are expected to perform for the project.

After authentication, the user will be presented with a list of options for data entry, data management, reporting, and other administrative functions. More information on the specifics functions and facilities of the web site are provided in the *T1DGC Manual of Operations*.

Data entry will be performed on a PC, with menu driven software developed in HTML for user interface without the requirement of site-specific software (other than Internet Explorer). This approach provides many attractive features for distributed data entry, including easy screen formatting to resemble paper forms, range and validity checking, easy retrieval, extensive reporting capability, and security.

2. Laboratory Data

Laboratory data will be transmitted to the Coordinating Center over a secure medium. Once received, the data will be imported into the SQL server repository. Online reports and monthly quality control reports are created to confirm that what was received matches what was sent.

3. Genetic Data

Genetic data will be transferred to the Coordinating Center electronically via a secure medium. Once received, it will be processed to verify the accuracy of the data. The data will be loaded into the SQL server data repository where it will be available for further processing.

4. Sample Inventory and Tracking

Use of bar codes for sample tracking and inventory are a critical component of the study. Samples will be collected at a variety of locations across four networks. Each Regional Network Center laboratory is responsible for entry of Shipping Forms received from clinics within their network. At the time of ascertainment, a bar-coded ID is assigned by the clinic staff.

The bar-coded labels placed on the aliquot tubes have been extensively tested and are guaranteed to remain intact and "scanable" under the most extreme of conditions. Because the labels are not "printed" on paper but "fused" in plastic, the bar codes are readable and nearly indestructible. Further, the label ends overlap when applied to sample tubes (with clear ends not to obstruct the label), forming a strong seal.

LITERATURE CITED

1. Cox NJ, Wapelhorst B, Morrison VA, Johnson L, Pinchuk L, Spielman RS, Todd JA, Concannon P. Seven regions of the genome show evidence of linkage to type 1 diabetes in a consensus analysis of 767 multiplex families. *Am J Hum Genet* 2001; 69: 820-830.