Manual 13
Data Management

September 28, 2014 - Version 2.0

Study website - http://www.cscc.unc.edu/hchs/

Data Management
TABLE OF CONTENTS

1. HCHS / SOL Field Center Data Management

The data management manual focuses on HCHS / SOL data collection and processing procedures at the field centers, at the central agencies and at the coordinating center. Topics include identification of participants, data collection forms and procedures, instructions for completion of paper and electronic forms (including corrections), data transfer and naming conventions, revisions to manuals and forms, and general guidelines to ensure data security, accuracy, and accessibility at field centers and central agencies.

1.1. Web-Based Data Management

The HCHS / SOL Data Management System runs on the second generation of the Carolina Data Acquisition and Reporting Tool (CDART) software. CDART is a set of customized data entry screens and report programs which the field centers and reading centers use to manage data collected in the HCHS / SOL study for the second examination. CDART runs over the Internet using web browsers such as Chrome or Firefox. The system will operate in from any computer with a high speed internet connection and a web browser. By design the CDART supports data entry either interactively "on screen" during data collection or from paper forms. However with the exceptions noted below, the data collection protocol specifies direct entry as the preferred mode of data collection. Work station configuration at the field centers should be arranged accordingly. In all circumstances when research data is collected on paper (hardcopy), key the data into CDART before the participant leaves the field center to permit queries, clarifications, and collection of missing items. The forms that may be collected on paper first are:

- Eligibility Screening and Scheduling (ELE)
- Screening Call Tracking (SCT)
- Exam Itinerary and Checklist (CHK)
- Informed Consent Tracking (ICT)
- Phantom form (PHT)

The separate HCHS / SOL CDART2 Data Management System User Guide provides a description on how to use CDART. There is also a companion online self-guided tutorial for CDART2 software at the URL, http://www2.cscc.unc.edu/home/?loc=training

Usernames and passwords for the system are provided by the Coordinating Center (CC) in a secure and confidential manner to each site's project and data managers. The CC will add and administer new users, and delete former users from the CDART system since that is a server access administrative function. However, it is the responsibility of each hiring manager at the HCHS field centers and reading centers to train their staff on the security, confidentiality, and ethical collection of research data and compliance with all local institutional and Federal contract information technology rules of behavior.

In summary, CDART provides several major functions that are fundamental for management:
Data Entry: Allows data collection forms to be keyed, edited and updated, either through the HCHS / SOL internet based data entry using CDART.
Reports: Provides customized reporting based on study need. CDART will generate participant recruitment reports, query lists at item and form level, form inventories, to help the field centers

with any data management-related tasks. Requests for reports or lists not provided by CDART can be made by the field centers to the CC. Participant recruitment, scheduling, and other managing of participant flow is the responsibility of the field center

Data Transfer & Uploading: Allows laboratory or reading center data to be sent to the HCHS / SOL Coordinating Center. This external data can be uploaded into the HCHS / SOL web server database. Data records process in batch via 'remote upload' are stored for pre-processing before being loaded by scripts into the consolidated database overnight.

1.2. HCHS / SOL Participant ID Numbers

ID numbers for participants enrolled in the HCHS / SOL study were created for the baseline examination (Visit 1) and assigned to each field center by the Coordinating Center as part of the original recruitment process. Those same ID numbers for all cohort members will continue to be used for the duration of the study in the second examination (Visit 2) and also in the annual follow-up interview process.

HCHS / SOL participant ID's are 8 characters long with the following format:

Character 1: Site Identifier (B=Bronx, C=Chicago, M=Miami, S=San Diego)
Characters 2-7: Participant ID number
Character 8: Check Digit (used for input validation of ID)

1.3. Identification Information on Data Collection Forms

The primary mode of data collection in is always assumed to be direct capture using the CDART entry system for procedure screens and for questionnaires. However, if data must be collected on paper it is important to record the key field information carefully. The information that identifies each form for a participant as a unique record in the HCHS / SOL database is the key field information contained in the "header" box at the top of the first page on all forms (see example below). The following guidelines should be observed in filling out the "header" information located at the top of the first page on all forms.

HCHS/SOL Personal Identifiers Questionnaire

ID NUMBER:									Visit Number (Event)	02	Seq#	0	1	Form Code: IDE Version: 1 10/12/2014

1.3.2 Form Code and Version

The form code is unique for each data collection form in the HCHS / SOL study. Form codes and versions are preprinted on all forms. If form content changes during the study, those

changes will result the version being updated from the initial version, "1" to the next number in the alphabetic sequence (2, 3, 4, etc.).

It is the responsibility of the site to make certain it is using the appropriate version of each form at all times. Note that CDART will load the current version of a form automatically. For details of the documentation procedures that are designed to facilitate communication to sites about form changes, see Section 1.9.

1.3.3 Visit

The Visit number (called Event in CDART nomenclature) will be used to differentiate a participant 'contact occasion' or visit over time. The visit occasion is pre-recorded on all forms for the examination. At present, all interview forms and procedures are collected at Visit "02" to since this exam is the second time point for HCHS / SOL full exam battery data collection. As the study participant follow-up schedule is defined over time, contact occasion is incremented designating follow-up year contacts.

1.3.4 Occurence (Sequencing Number)

The reoccurrence within a visit of forms that can be repeated use a sequence number (Seq #) to note the next occasion which enumerates and distinguishes multiple forms collected during a same visit number. For all forms at the regularly scheduled contact the occasion number is simply, "01". Where appropriate, the sequencing number has been pre-coded on the form. Any forms that are repeated for use in quality control purposes will have a different sequence number (such as "02", "03") to distinguish them from regularly scheduled visit forms.

1.4. Administrative Information

1.4.1 Code Number IDs

Certified site staff will be assigned a 3-digit code number by the Coordinating Center. The staff member must use this number on data collection screens that ask for a staff code number ID.

1.4.2 Date

The date to be recorded onto a data collection form header is the date of the participant contact or specimen collection (i.e. clinic visit examination date), or the date the form is completed. The date must conform to the month / day / year format as specified and be within the bounds of the time line. Pre-dating and post-dating of forms should not be done.

1.5. Data Collection

1.5.1 Background

HCHS / SOL uses a combination of data collection methods: direct data entry, recording on paper forms followed by data entry, and forms collected on paper only (with no data entry). The

purpose of this section is to provide instructions for completing forms. Prior to working with the forms, both this section and the specific question-by-question set of instructions for each form (QxQ) should be read carefully. (The QxQ instructions will follow the paper form in the MOP.)

1.5.2 Form Structure

The HCHS / SOL DMS Data Entry screens are designed to parallel the paper forms on which the data are collected. The general layout of the paper forms is as follows:

First Page of Form:
Form Title
"Header" Information
Participant's ID Number
Form Code and Version
Visit (Event or Contact)
Occasion (Sequence Number)
Data collection questions

In the CDART, the header information is entered into the ID screen and subsequently displayed in the header frame of the forms grid and at the top of the form during entry of data items.

1.5.3 Recording Responses to Questions

Most of the questions in the HCHS / SOL questionnaire battery of forms have a set of pre-coded responses, with instructions to “enter the appropriate response” (code) or “check all that apply” (checkbox), or selection of a listed option. However, a few questions require a written response. Others request elaboration of an “other” or “specify” response from a previous question. Space is provided on the form for those unstructured written responses.

If a participant’s answer does not logically fit into one of the pre-coded answers, the interviewer must specify the response by recording it on the form beside the pre-coded answers. Data entry personnel are trained to enter the additional data into note logs or to use the status field to indicate that the participant does not know the answer, or even refuses to answer the question.

The data collection practices below must be followed at all times to assure that the recorded response accurately reflects the participant’s answers and that questionnaire data can be converted to a computer-readable format.

Guidelines for the interviewer include:
Carefully listen to what the participant says and record the appropriate answer if the response satisfies the objective of the question.

In recording answers to open-ended questions or “other” categories, record the response in the participant’s exact words.

On paper, record in the white space (below the questions) any responses "that don't quite fit" in one of the response categories. The interviewer's notes will help the data analyst to understand points of confusion, difficulty, etc. Notes on paper forms can be entered as "note logs" in the DMS.
On paper, print or write legibly.

- If a participant refuses to answer a question, and "refused" is not a value in the response set, write "refused" in the left-hand margin beside the question and enter **equal signs** ("=") in the response field to signify a double strikethrough.
- A single answer code must be circled / entered for each question to represent the participant's answer.
- A "select all that apply" answer pattern is indicated with a checkbox, or with instructions to "circle all responses that apply".

Some of the questions in the HCHS / SOL study ask about recall of events over time. The interviewer may assist the participant without violating probing rules by working with him/her on converting dates to duration (e.g. "for how long did you….") or pinpointing dates or events. Another way to help with the collection of more accurate information is to ask the participant to think about the time of year or season when an event occurred.

1.5.4 General Instructions for Completing and Correcting Items on the Forms

General guidelines for the interviewer regarding forms:

Review each form and its instructions prior to use. If you are collecting data on a paper form prior to data entry, verify that you are using the appropriate form by checking its 3 letter form code, version, and date, all located in the lower left-hand corner of the page. Each unique form type will have specific instructions for filling out that form in the Procedures Manual. Be familiar with the instructions in the Procedures Manual before attempting to complete a form. Print all text responses legibly; do not use cursive writing if collecting data on paper first.

All items fall into one of three main categories: (1) "fill ins", (2) multiple-choice (circle or check), and (3) qualitative information (comments/short-answer questions). Techniques for completing each of these types of items, as well as making corrections, are described below. A general rule is to record information only in the spaces provided (except for some error corrections).

- Corrections to paper forms should be made in the following manner:
- Cross out the original response with an 'X' in such a way that it is still legible.
- Write the correct response above or to the side of the original response.
- Date and initial the correct response.
- In cases where repeated corrections were made to the same response, circle the final corrected response.
- Major changes should be documented with a brief explanation in the margin.

Corrections to electronic forms are made using the DMS. The DMS records the date and time of the update and the user who makes the change. Many corrections will be made in response to queries sent to the field center from the CC.

Do not attempt to correct errors by using correction fluid or erasers at any time. Data collection forms need to maintain the history of data recorded in the event of an audit. The audit log in the DMS maintains this history for forms which have been entered and subsequently corrected (but does not track paper-only corrections).

When a response is outside the normal limits or seems contradictory based on other data, confirm the data and, if using paper, write "confirmed" in the margin. This will decrease time-consuming queries and help the data entry staff.

Carefully proofread each page of data for legibility, accuracy, and completeness prior to transferring the form to the data entry staff.

1.5.5 "Fill Ins": Recording Information

"Fill Ins" refers to items where the question is given a defined space for recording the answer. Questions asking date, ID number, height, weight, etc. will have a limited amount of space for data entry. In the event that the response contains more characters than there is room for in the space provide by the form, indicate the correct response in the form margin near the original response, and enter the value into a note log in the DMS.

Numeric fields may have a preprinted number of decimal places. In this case, the QxQ instructions will specify the number of decimal places to be recorded. Instructions on how to round values to the expected number of decimal places are found in the QxQ instructions.

When a date is recorded, slashes ("/") are used as the separator characters for month, day, and year. These are usually preprinted in the response field on the paper form but must be entered into the DMS. The format to be used to record dates is indicated below the boxes. HCHS / SOL uses the U.S. order for recording dates (month/day/year).

HCHS / SOL usually records time using a 24-hour clock (noon= 12:00, 1pm = 13:00). In most cases, colons (":") are used as the separator character for hours and minutes and are typically preprinted in the response field and must be entered for questions where hours and minutes are not separate questions.

1.5.6 "Fill Ins": Correcting Mistakes on Paper Forms

If a number or letter is entered incorrectly, the person making the correction should first mark through the incorrect entry with an "X". Then, he/she should clearly code the correct entry above the original (incorrect) entry and initialize the correction, using his/her 3 initials, and record the date of the correction.

If a mistake is made and corrected, and then it is discovered that the correction is incorrect, make a second correction using the same rules as above.

1.5.7 "Fill Ins": Unknown or Inapplicable Information

If an item of this type (either alphabetic or numeric) does not apply to the participant being interviewed, leave it blank. For example, if the participant does not have an "other" phone number, that item is left blank. Similarly, if the form provides spaces for three measurements, but only two are taken, the third space is left blank.

1.5.8 Multiple-choice: Recording Information

In this type of question several alternatives are given for the answer, each having a corresponding letter and/or word. When it is decided which alternative is most appropriate, circle the corresponding letter on paper. Always circle one letter/word only. Key the letter or word into the DMS when entering the data.

1.5.9 Multiple-choice: Correcting Mistakes on Paper Forms

If a response is coded incorrectly, mark through the incorrectly coded response with an "X" and circle the correct response. Initialize and date the correction.

1.5.10 Qualitative Information: Recording Information

Some forms contain a substantial amount of qualitative data. These short answer and comment questions will be handled differently than "fill in" type data. If these types of questions are filled out by hand on paper forms, write clearly. When keying these responses, any unreadable answers/comments should be answered to the best of the data entry personnel's ability with a note describing what was illegible.

1.5.11 Skip Patterns

Generally, questions are answered in the order presented, with none omitted.

A skip pattern directs that one or more questions be omitted (skipped) when they are not pertinent to the participant's situation. (For example, if question 1 pertains to sex of the participant, and question 2 references recent or past pregnancies, then a male respondent would be directed to skip question 2.)

Skip patterns occur in some multiple-choice items. This may be indicated on the form by an "If _____, go to question #" statement. If response associated with the skip is selected, the next item to be asked is the one indicated in the "go to" statement.

If a skip pattern response is not selected, proceed to the next item in sequence as usual. Occasionally, a skip pattern will occur in a "fill-in" item (such as, "Other, specify"). If the skip criteria are not met, continue to the next item as usual.

1.5.12 Problem Clarification and Data Queries

CDART is programmed to automatically query out-of-range values during the data entry process. However there may be a need to send queries from the CC regarding data values within or across forms. All queries will be sent electronically to the data coordinator, participants will be identified by ID number, forms will be identified by header information, data items will be identified by question numbers, the original response will be indicated, and the reason for the query will be described. A cover memo will accompany the data queries describing the problem, with suggestions of ways to resolve the problem and a timeline.

1.5.13 Permanently Missing Forms

In the event that a participant is unable to complete an exam, all forms for the contact (or visit) which were not completed should be entered into CDART with the "Permanently Missing" status.

1.6. Security of Paper Forms

Each clinical site is responsible for assuring that participant study data is stored in a secure location that meets participant confidentiality requirements.

1.7. Data Management Reporting

The DMS has numerous reporting programs to facilitate data management at the sites. The HCHS/SOL CDART2 User Guide contains the documentation of the reports available in the DMS. Information on updates and changes to these reports will be provided through the Numbered Memo communication (Section 13.12.1) from the Coordinating Center to the field centers. As these DMS reports are updated or changed, training conference calls with the field center data coordinators or project managers may also be scheduled.

1.8. DMS Training and Certification

The web based training seminar hosted in Chapel Hill in August 2014, before Visit 2 will assume that each trainee has viewed the online self-guided tutorial at the following link.

http://www2.cscc.unc.edu/home/?loc=training

All users of the CDART system at a SOL study site should attend the central training webinar and complete the certification session for CDART. Those attending the central training may provide additional training to other staff members at their sites. Follow-up conference call training sessions will be scheduled as needed. Monitoring site visits by Coordinating Center personnel are scheduled to take place throughout the study; some DMS training can occur during these visits as well (depending on the circumstances and the perceived need for such training).

1.9. Official Study Documents

Current versions of all study documents, protocol, data collection forms, MOP, user's guides, and other important documents are available on the study website at http://www.cscc.unc.edu/hchs. They are in the Study Members' area under the "Documents → Protocols and Manuals", and "Training and Information Materials" pages. To access them the

user must supply a username and password. Each document exists as a PDF file. It is recommended that the PDF files be printed at the sites because formatting and special characters are retained. MS Word files for informed consent templates are kept on the web site to facilitate working drafts as needed.

IMPORTANT: Versions of these documents that are designated as usable in the field (otherwise called "final") were sent to each study site as hard copy in an official HCHS / SOL Study Documents Notebook. One notebook will be provided to each Project Coordinator and Steering Committee Member. Section 13.8.1 describes the process of communicating updates to documents and of verifying receipt of communication on modifications.

1.9.2 Instructions for New / Corrected Materials

Forms: Any new or corrected forms will be available to print from the website. Forms should be replaced and copied for immediate use. Email confirmation must be sent to the CC (HCHSAdminstration@mail.cscc.unc.edu) when the revised forms are downloaded from the Internet.

Manual: The revised pages/chapters of the HCHS / SOL Manual of Procedures should be printed from the website and filed immediately in the MOP notebook. Email confirmation must be sent to the CC (HCHSAdminstration@mail.cscc.unc.edu) when the revised pages/chapters are downloaded.

QxQ's: Any new or corrected QXQ will be available to print from the website. They should be printed and filed immediately together with the appropriate form in the MOP binder. Email confirmation should be sent to the CC (HCHSAdminstration@mail.cscc.unc.edu) when the new QxQ's are downloaded.

1.9.3 Instructions for Outdated Materials

All outdated pages of the MOP, forms or QxQ's should be removed from the Documents Notebook as instructed in the Numbered Memo. Outdated materials should be archived according to each site's institutional requirements. All study materials are archived at the CC.

1.9.4 General Filing Instructions

All participants should have either a binder or file folder filed in alphabetical order by participant ID. If the center prefers to file by last name, there should be a cross-referenced list available with the corresponding ID number. It is important for centers to be able to communicate effectively with the CC by the participant's ID number. Data queries sent to the sites from the CC will only identify a participant by ID number. Remember, before sending any forms to the CC; blind (or mask) all personal information pertaining to the participant.

The safety and confidentiality of the study data and equipment is the responsibility of each study site. All computers, memory keys, diskettes, and participant data must be stored in a secure location.

2. CENTRAL AGENCY DATA MANAGEMENT

Data management at each of the six reading centers will vary. This section describes the practices at each agency and the standards for data transfer established for HCHS / SOL.

2.1. Echocardiography Reading Center

The HCHS/SOL cohort re-examination includes an assessment of cardiac structure and function in participants of the age of 45 or over using echocardiography (a cardiac ultrasound). Echocardiograms are performed according to a common protocol by HCHS/SOL sonographers centrally trained and certified by the Echocardiography Reading Center, and images are analyzed at the Echocardiography Reading Center at the Brigham and Women's Hospital Cardiac Imaging Core Laboratory. The scanning and reading procedures, as well as the training and certification of the sonographers and readers are detailed in protocol manuals 17a and 17b, respectively. Participants who had an echocardiography examination through the "ECHO-SOL" ancillary study will not have one performed at Visit 2 of the SOL. Each month the reading center will upload files using the HCHS/SOL CDART web portal site to the CC for inclusion in the study database.

2.2. Central Laboratory

The University of Minnesota, Advanced Diagnostic Research Laboratory will receive, store and analyze the urine and blood chemistries for the specimens transferred from the field centers. The central laboratory will upload results from batch runs performed weekly and provide that data on at least a monthly basis to the study sites by making it available through the DMS interface. Each transfer from the central laboratory to the CSCC will comprise a single Excel spreadsheet. These worksheet names should not change without notice. Variables for each data set are defined in the first line of the sheet and are different for each set of data. The names in the spreadsheet are mapped to standard DMS variable names through a table called LABNAMES. It is important that once these names are agreed upon, they do not change without notice. The files are named according to study, lab and transfer sequence.

Central laboratory analyses are sent incrementally as samples are analyzed by the University of Minnesota Fairview Hospital Clinical Laboratory. If a record needs corrections to result values, it is included in a later transfer. The new record overwrites the original one.

The Excel files are uploaded to the CSCC using the HCHS /SOL DMS web site. Files are processed into the HCHS / SOL consolidated database overnight and available for reports the following day. Clinic visit results are added to the clinical laboratory (LAB) data table. The variable mapping is described in the data transfer reference manual for labs and reading centers.

Table 1. HCHS/SOL Central Laboratory Reporting Reference and Alert Ranges

Test Name	Reference Range	Units	Reported	Alert Value	Sample Type
Hemogram (CBC):					
White Blood Count (WBC)	4.0-11.0	x 10^9/L		<2 and >25	EDTA - whole blood
Red Blood Count (RBC), male	4.4-5.9	x 10^{12}/L			EDTA - whole blood
Red Blood Count (RBC), fem.	3.8-5.2	x 10^{12}/L			EDTA - whole blood
Hemoglobin - male	13.3-17.7	g/dL		<8	EDTA - whole blood
Hemoglobin - female	11.7-15.7	g/dL		<8	EDTA - whole blood
Hematocrit - male	40.0-53.0	%			EDTA - whole blood
Hematocrit - female	35.0-47.0	%	Yes		EDTA - whole blood
Mean Corpuscular Volume (MCV)	78-100	fL			EDTA - whole blood
Mean Corpuscular Hemoglobin (MCH)	26.5-33.0	pg			EDTA - whole blood
Mean Corpuscular Hemoglobin Concentration (MCHC)	32-36	g/dL			EDTA - whole blood
Red Cell Distribution Width (RDW)	10.0-15.0	%			EDTA - whole blood
Platelet Count	150-450	x 10^9/L	Yes	<50 and >1000	EDTA - whole blood
WBC Differential:					
Neutrophils	40-75	%			EDTA - whole blood
Lymphocytes	20-48	%			EDTA - whole blood
Monocytes	0-12	%			EDTA - whole blood
Eosinophils	0-6	%			EDTA - whole blood
Basophils	0-2	%			EDTA - whole blood
Absolute Neutrophils	.6-8.3	x 10^9/L			EDTA - whole blood
Absolute Lymphocytes	0.8-5.3	x 10^9/L	No		EDTA - whole blood
Absolute Monocytes	0-1.3	x 10^9/L			EDTA - whole blood
Absolute Eosinophils	0-0.7	x 10^9/L			EDTA - whole blood
Absolute Basophils	0-0.2	x 10^9/L			EDTA - whole blood
Total cholesterol	0-200	mg/dL	Yes		Serum
Triglycerides	0-150	mg/dL	Yes	>1000	Serum
HDL-cholesterol	#	mg/dL	Yes		Serum
LDL-cholesterol	#	mg/dL	Yes		Calculation
Glucose, fasting	60-115	mg/dL	Yes	<60 and >200	EDTA - plasma
Glucose, post OGTT	0-139	mg/dL	Yes		EDTA - plasma
Glycosylated Hemoglobin	4.3-6.0	%	Yes		EDTA - whole blood
Insulin, fasting	2-25	mU/L	No		EDTA - plasma
Insulin, post OGTT		mU/L	No		EDTA - plasma
Creatinine, male 20-50y	0.8-1.5	mg/dL	Alerts only	>2.0	Serum
Creatinine, male 50-60y	0.8-1.7	mg/dL	Alerts only	>2.0	Serum
Creatinine, male >60y	0.8-2.0	mg/dL	Alerts only	>2.0	Serum
Creatinine, female 20-50y	0.6-1.3	mg/dL	Alerts only	>2.0	Serum
Creatinine, female 50-60y	0.6-1.4	mg/dL	Alerts only	>2.0	Serum
Creatinine, female >60y	0.6-1.6	mg/dL	Alerts only	>2.0	Serum
eGFR	<60, >60	mL/min/1.73m^2	Yes		Calculation
Albumin/creatinine ratio	0-20	mg/g creatinine	Yes		urine, neutral
Creatinine, urine	NA	NA	NA		urine, neutral

3. HCHS / SOL COORDINATING CENTER DATA MANAGEMENT

Computers at the UNC CSCC are connected via a Local Area Network. The network includes clustered database servers running the Novell network operating system and is connected to a Storage Area Network (SAN). Clustered servers running the Microsoft network operating system provide web services for the data management system. The web servers are isolated by a router from the servers holding study data.

The consolidated database will be stored in a SQL-server database. Standard transaction validity checks will be applied to all updates to the database (e.g., to prevent the addition of records with duplicate keys, etc.). Audit logs from the DMS provide complete documentation for changes to the consolidated database. Backups of the consolidated database as well as processing reports are made daily. Once a month, the current backup tape is permanently archived at an off-site data storage facility. Periodically the consolidated database goes through a series of closure checks to ensure the completeness and correctness of data collection and processing. These checks are performed on a 'frozen' version of the database defined by a specific time cut point. Typical closure checks include classifying the universe of IDs, assuring all expected forms were received and assuring all queries generated were resolved.

3.1. Central Agency Data

Data from central agencies will be processed into the consolidated database every night or the night after receipt (if not entered via web-based DMS). An automated process will read the data, perform key field and record level integrity checks and add valid records to the database. Data from each agency is processed and results reports are generated. The report lists key fields of records which were not processed due to error. It also lists total number of records processed without error. The processing reports are available to the agencies, the field centers and CSCC HCHS / SOL staff.

3.2. Field Center Data

Field center data, entered via the web-based data management system, will be copied from the web server to the local area network (LAN) every night. From the files on the LAN, data will be retrieved into statistical analysis files for use in study reports for the steering committee and NHLBI.

3.3. Reports

The DMS will provide each field center with the ability to generate a variety of reports. These include participants contacted and examined, indicators of data quality, completion status of participant result reports and specimen tracking reports, among others. Such reports make it easy for study coordinators to monitor their center's performance and the timely identification and resolution of problems in data collection. The reports for study participants and their dentists or physicians are described in Section 3.8, Reporting of Study Results.

The field center data quality reports are complementary to the monthly Steering Committee reports. The former can be run in real time by field center staff and access up-to-date data stored in the consolidated data base. The Steering Committee reports will be produced monthly (or a schedule defined by the Steering Committee) and thus reflect the status of the study at the time of the most recent retrieval. Their purpose is to provide the Steering Committee and center investigators with performance information at all sites.

3.4. Data Security and Confidentiality

Data confidentiality and security will be applied at all levels of data acquisition, transfer and storage, and applied to all study agencies, from field centers to coordinating center. The DMS developed by the UNC CSCC meets exacting data management standards of confidentiality, as well as HIPAA requirements. Beyond the password controlled access to the study equipment and the DMS, data collected at the field centers and in hospital record rooms are encrypted by the system and can only be decrypted for display on-screen by authorized study personnel. Personal identifiers are collected on separate forms (and transferred as separate, encrypted records). The Coordinating Center will be responsive to data confidentiality requirements originating from providers of medical care or IRBs, as needed to enable the work of the field centers. When paper data collection forms are used they will be retained at secure locations at the field centers until the Steering Committee acts on recommendations from the Coordinating Center to dispose of such records (e.g., incremental data closure). The secure storage and disposition of hard copy records at field centers will follow institutional procedures at each site.

The DMS server will be housed at the Coordinating Center and exclusively managed by CSCC personnel. Measures to ensure the security of the data include: restricting access to users with valid IDs and passwords; using a firewall to restrict access to the web server and to shield the UNC CSCC LAN from web users; using the secure sockets layer standard to provide encryption and user authentication. In accordance with CSCC standard operating procedures, system security logs and event logs are monitored daily to detect unauthorized attempts to access the system. The UNC Information Technology Systems group publishes a guide called “ITS Security at UNC Chapel Hill – Securing IIS”. The UNC CSCC follows these guidelines, which include closing unused ports; requiring user passwords to be long and difficult to guess; deleting certain files and subdirectories; and managing file and folder privileges.

All data transferred to the Coordinating Center will be stored, processed, and analyzed within the UNC CSCC office suite. At the CSCC, all access to office space containing data is controlled through locked doors. Visitors may enter only when accompanied by a UNC CSCC escort. All office space remains locked after working hours. Access to computer data files is controlled by passwords released only to the UNC CSCC personnel who use such files. In addition, data files with personal identifiers (and sensitive information per designation by a study’s Steering Committee) are encrypted.

As standard practice, output mailed to a field center identifies participants only by ID number. No individually identifiable information will be distributed by the Coordinating Center to any study agency other than the originating field center. Printed material containing confidential information is discarded through supervised loading, transportation, and storage using a chain of custody control process, until the material can be recycled into paper pulp.

It is a requirement for all CSCC staff to complete a confidentiality certification procedure upon employment. Policies regarding the confidential nature of the data collected, processed, and stored at the UNC CSCC are explained to all personnel, who must then sign a "confidentiality certification" to be allowed access to confidential information. The CSCC reinforces the confidential nature of all study data at its staff meetings.

3.5. Data Retrieval and Statistical Computing

Data will be retrieved from the study database and converted into SAS files on a regular schedule (e.g., monthly). The retrieved files will be stored as SAS datasets within a SAS data library. Most statistical computing will be done using SAS software. All statistical computing will be performed by a dedicated statistical programming staff, using a well-established statistical computing request system that has proven itself through use with many long-term, multi-center research projects managed by the CSCC. This system includes thorough documentation of requested computing, programming standards, naming conventions for datasets, programs and program results, inventorying and tracking of computing requests, procedures for program review, and permanent archival of completed programs, results, and datasets.

3.6. Database Closure

Data queries will be generated on a monthly basis, immediately following data retrieval. Typical data checks include classifying the universe of enrolled IDs, assuring all expected forms were received, performing consistency checks between related data fields, assuring all queries generated are resolved, etc. If there are unexpectedly high error rates for a site or a user, we explore the causes of the error and take corrective action, such as retraining personnel or making changes to the data management system. Research indicates that this comprehensive data checking, in combination with extensive real-time edits, can substitute for double data entry for data entered from paper forms.

Periodically the study's consolidated database is subjected to closure checks for completeness and accuracy of data collection and processing. These checks are performed on a "frozen" version of the database defined by a specific time cut point, and precede the use of data for publication. Typical closure checks include classifying the universe of IDs, assuring that all expected forms were received and all queries were resolved, examining the consistency of items across forms and visits, and checking distributions of key variables for possible errors. Current plans entail closing the database in waves, one per examination year so that investigators will have access to interim results for study monitoring, review, and publication.