SEES Data Management Plan

1 Data Management

The purpose of this section is to describe the data management processes, both manual and electronic, implemented for the SEES project. All data will be collected electronically through web-based surveys; there will be no paper based survey data. In order for the data to be collected securely, the process is not only dependent on controls, but on software and hardware to both allow data collection and enforce those controls.

1.1 Software and Hardware Requirements

This section lists the software and hardware required for successful data management throughout the SEES project.

1.1.1 Software Requirements

The following software will be used in data collection and analysis:

1. **IBM SPSS Data Collection 6.0.1** to program the web-based survey questions and electronically record interview responses in the SPSS SEES project database as they are submitted through the Safari browser on the iPads.

2. **SQL Server 2008** to store the data transferred on a nightly basis from SPSS. This data will be in a format that is directly exportable to SAS for analysis.

3. **Bitvise** to provide the secure connection that the data will be transferred over from SPSS to SQL Server.

4. **SAS version 9.2** to analyze the interview data collected with SPSS.

The following software will be provided by CDC:

1. **SDN Digital Certificate** to transmit data files from SciMetrika to CDC via the Secure Data Network (SDN). A digital certificate is required to use the SDN.

1.1.2 Hardware Requirements

The following hardware will be used in completing the SEES data collection and management:

1. **iPads** to collect data from field administered web-based surveys. No information will be stored directly on the iPads; instead, the information will be transmitted securely through the web-based survey to the SPSS SEES project database.

1.2 Data Configuration

To ensure consistency in database layouts, CDC will provide the survey specifications. These specifications will cover the questions, variable names, field limits, consistency checks, response values, and formats.

1.3 Databases and Data files

The following 2 databases will be used to store the data, and ultimately create any necessary data files:

1. SPSS SEES project database for the clinic and community surveys
2. SQL Server database for the formatted data from the clinic and community surveys.

In addition, the final normalized and de-normalized datasets will be exported in SAS to CDC.

1.4 Data Submissions to CDC

The raw data collected with the web-based questionnaire will not be sent to CDC; however, the normalized and de-normalized SAS datasets for each of the 6 jurisdictions for all 4 outcome evaluation data collections will be sent to CDC on or before the end of the contract in September 2015. The 4 outcome evaluation data collections will consist of the 2 clinic based Outcome Surveys and the 2 community based Outcome Surveys.

These data sets will be submitted to the CDC data manager via the SDN.

1.5 Quality Control

1.5.1 SPSS Survey Data

The web-based survey collection that implements rigorous logic to ensure data integrity, and as a result, the data management needs are expected to be minimal. However, to ensure further data integrity, the site coordinator for each jurisdiction will execute a web-based tool that validates the data for completed surveys, reporting any responses that are not valid or within range as defined in the survey requirements supplied by CDC. This validation is run against the data stored in the SPSS SEES project database that was obtained from administering surveys in the field. If an issue should come up with a survey during this validation, we will be able to identify it early and correct it. These validations will be run during at the end of each day of data collection.

Similarly, the staff member designated as the data manager for the SQL Server database at SciMetrika will run these validation routines to ensure that the data transferred from the SPSS SEES project database meets the same data integrity. This process will be run within 24 hours after each data transfer.

In addition to validation of the survey data for quality assurance, the site coordinator at each site and the SciMetrika data manager will develop a system for tracking data problems and changes to the SPSS SEES project database.

Finally, an automated data cleansing process will be applied to clean the data. This cleaning process will clear variables that were entered during the administration of the survey but need to be cleared due to the respondent needing to go back and change answers that resulted in a different logic flow. The process will be based on the logic defined in the survey instrument supplied by CDC, and the raw data will not be altered. Instead the process will create a new “cleaned” data set that will be stored in the SQL Server database. This is the dataset that will be exported to SAS for analysis.

1.5.2 Interviewer Monitoring

As taken from the Data Collection plan, the following interviewer monitoring will be implemented:

All interview data are vulnerable to bias from variability in the way the interviews are conducted. This bias may arise from variability between interviewers or from variability between interviews conducted by a single interviewer. To prevent these biases, and to ensure that proper procedures are followed, monitoring procedures will be implemented to assess the consistency and quality of interviewing and the quality of data collected.

Monitoring begins with the training of the interviewers as role-play with the surveys is required during training, and their performance is monitored to ensure training success.

After interviewers have successfully completed SEES training, the SEES site coordinator will regularly monitor each interviewer as they conduct the eligibility screener, obtain informed consent, and administer the survey questionnaire during the data collection process. Feedback on the interviewers’ performance – areas of proficiency as well as areas for improvement – will be discussed with them shortly after observations are conducted. Interviewer evaluation forms will also be shared with SciMetrika project staff for additional monitoring.

1.5.3 Data Editing

As defined in 1.5.1 above, there should not be a need for data editing; however, should data editing be required, the site coordinator at each site and the SciMetrika data manager will develop a system for tracking data problems and changes to the SPSS SEES project database. Currently, the process will require that the Survey ID be used to identify the survey in question, then the data in question will be submitted through the site coordinator to the SciMetrika data manager to determine the necessary corrective action.

1.6 Backing Up Data

Backing up data minimizes the loss of data. The loss of data can occur due to database corruption, including deletion, hard drive failure, theft of equipment, and “catastrophic” events. To prevent against the loss of data, backups are stored both locally and at a secure remote location on a regularly scheduled basis. This backup policy is detailed in the SEES System Security Plan.

1.7 Computer Virus Protection

As detailed in the SEES System Security Plan, all systems are protected by Symantec Anti-Virus malicious code protection mechanisms at information system entry and exit points and at workstations, servers, or mobile computing devices on the network to detect and eradicate malicious code. Detailed information regarding computer virus protection can be found in the SEES System Security Plan.