3. SYSTEMS

This chapter describes the NICLTS computer systems and related activities needed for data collection, data transmission, key entry, data cleaning, and data summarization. Since these activities differed between the two major phases of the study, the discussion is presented in two parts:

- Phase I--Collection and processing of data from laboratories with moderate and high complexity CLIA certificates and

- Phase II--Collection and processing of data from laboratories with waived and PPM CLIA certificates.

3.1 Phase I-Collection and Processing of Data from Laboratories with Moderate and High Complexity CLIA Certificates

Data collection in laboratories with moderate and high complexity CLIA certificates was performed by medical technologists reviewing hard-copy laboratory records in clinical laboratories. These technologists, who were referred to as "tabulators," recorded the data using a laptop computer running a PARADOX for Windows program called "the Tabulation Device." The data were then transmitted to Westat's home office where they were converted to SAS files, cleaned, and summarized.

The following discussion is divided into two sections:

- Tabulation Device and

- Data Processing.

These sections correspond to activities taking place in the field and those taking place at Westat's Rockville offices.

3.1.1 Tabulation Device

The Tabulation Device allowed the tabulators to record identified clusters of triples consisting of an analyte, a test system, and a biological specimen that represented tests performed in a clinical laboratory and to

record the number of each cluster performed during calendar year 1996 at the laboratory. The Tabulation Device contained numerous features designed to protect the confidentiality of the data, while ensuring that it was collected as completely and consistently as possible.

3.1.1.1 User Identification

Each user of the Tabulation Device was assigned a computer-generated password consisting of a randomly selected collection of eight numbers and letters. This provided 1.7656×10^9 different possible passwords. Fewer than 50 of these passwords were assigned. Whenever a tabulator started the Tabulation Device software, he or she was required to enter this password. Only NICLTS IDs[4] of laboratory locations assigned to the tabulator possessing that password and currently in a "fielded" status were then active on the computer. Thus, a tabulator could enter data only for laboratories assigned to him or her. Each cluster (analyte, test system, biological specimen) that a tabulator recorded contained a unique identifier indicating who collected the data.

Each computer was also given a unique computer hardware identification number and was then assigned to a specific tabulator. Computer assignments were carefully tracked whenever a tabulator received a computer or returned one for diagnosis, repair, or closeout.

3.1.1.2 Laboratory Identification

In order to confirm that a laboratory was in fact the sampled laboratory, the Tabulation Device displayed the CLIA ID number of a laboratory after the tabulator logged on and selected it for data entry. This allowed the tabulator to compare the sampled CLIA ID number with the number on the CLIA certificate at the laboratory.

To ensure the confidentiality of the data, all assignments and data records used the NICLTS ID rather than the CLIA ID number. In order to preserve this confidentiality while making the CLIA ID number available to the tabulator when needed, the Tabulation Device contained an encrypted file that matched the NICLTS ID to the CLIA ID number. This file could be read only by someone using a valid password. The Tabulation Device was

[4] Each laboratory was given an arbitrary five-digit NICLTS identifier for use during the study.

designed to display the CLIA ID number only for the laboratory currently being worked.

3.1.1.3 Computer-Assisted Data Entry

The tabulators used the Tabulation Device to assign each identified laboratory test to an explicit cluster (analyte, test system, biological specimen); to record the calendar year 1996 volume of each cluster; and to record comments about the tests. Each full data record for a laboratory test consisted of the following:

- NICLTS location ID (five-digit NICLTS ID + two-digit location number) automatically inserted by the system;
- Site number (two digits) - automatically added by the system;
- Tabulator ID - automatically added by the system;
- Tabulation date - automatically added by the system;
- Analyte name - entered by tabulator (computer assisted);
- Analyte code 5 - automatically added by the system based on the tabulator's choice of analyte name from a lookup table (or left blank if no entry existed);
- Test system name - entered by tabulator (computer assisted);
- Test system code[5] - automatically added by the system based on the tabulator's choice of test system name from a lookup table (or left blank if no entry existed);
- Biological specimen name - entered by tabulator (computer assisted);
- Biological specimen[5] - automatically added by system based on the tabulator's choice of biological specimen name from a lookup table (or left blank if no entry existed);
- Volume[5] - entered by tabulator;
- Nonpatient Care Volume - entered by the tabulator, included any quality control, proficiency testing, method comparison testing, repeat patient testing and calculated values that were included in the Volume entry;[6]

[5] The analyte code, test system code, biological specimen code, and volume are the core information collected by the tabulator.

[6] The Nonpatient care volume entry was subtracted from the volume entry during data processing so that only patient care volumes are reported in this study.

- Comments - entered by tabulator (optional);
- Beginning and ending dates for the period represented by the entered volume automatically set to the beginning and end of the period being tabulated, usually 1/1/96 and 12/31/96, but could be edited by the tabulator if necessary; and
- New pair flag - automatically set to "I" if the analyte code and test system code pair did not exist in the expanded Complexity Model.

The Tabulation Device contained a number of features to assist the tabulators:

- Complexity Model;
- Data validation checks;
- Daily log sampling; and
- Nursing station sampling.

Each of these features is described separately below.

CDC Complexity Model

A database of analytes and test systems became the basis for data collection and was an integral component of the Tabulation Device. The database consisted of the CDC Complexity Model and additional analytes and test systems identified by Westat. The Complexity Model is a database of analytes and test systems linked to each other. Analyte is defined in the CLIA 88 regulations published in the Federal Register of February 28, 1992 (p.7139) as "a substance or constituent for which the laboratory conducts testing." An example is glucose. A test system measures the analyte present in a biological specimen derived from a human being, according to the regulations, for the purpose of "providing information for the diagnosis, prevention or treatment of any disease or impairment of, or the assessment of the health of human beings."

Since specimen type helps to determine complexity and since specimen is sometimes confounded with analyte and test system in the Complexity Model, Westat and CDC agreed to include specimen type as part of the data collected. A total of 42 unique specimen types were identified and given two-digit codes. The data tabulator's primary task was to identify each cluster comprising an analyte, a test system, and a specimen in a laboratory and to record each cluster's calendar year 1996 test volume.

The CDC Complexity Model contains thousands of combinations of analytes and test systems, yet a study of the database revealed that many items were missing, particularly in the microbiology specialties. For example, although "Growth or No Growth of Bacteria on Solid Culture Media" was included, parallel entries did not exist for fungi or mycobacteria. or for organisms ruled out on certain selective and differential media. Westat needed a means for tabulating all growth or no growth tests. Other examples of entries missing from the database included various methods for performing antibiotic susceptibility tests such as manual minimal inhibitory concentrations, antifungal susceptibilities, serum killing levels, and synergy studies.

There were also no entries in the CDC database for tabulating tests in the Anatomic Pathology specialty. Westat developed a general approach to classifying information in the areas of Cytology, Histopathology, and Chromosome Analysis using reference laboratory service manuals from laboratories all over the country; textbooks in many of the specialties; expertise from the medical technologists assigned to the project; and nonproprietary databases.

Westat added 571 analytes and 763 test systems to the list used while collecting the inventory. These entries were assigned temporary Westat placeholder codes until CDC could study them and assign permanent codes. The CDC database and the additional entries identified by Westat became known as the expanded Complexity Model. This expanded database became an integral part of the Tabulation Device.

The expanded Complexity Model was used as a means of editing the data entered by the tabulators. Whenever a tabulator entered a cluster that did not exist in the expanded Complexity Model, the system warned of a potential data entry error. Since tabulators sometimes needed to enter clusters not in the expanded Complexity Model, the system allowed them to continue but set a "new pair" flag in the database so that the record could be examined and, if necessary, edited at Westat's home office. Tabulators were asked to complete a comment field, providing any relevant information, whenever they encountered such a cluster.

In some cases, as discussed later, home office editors found that the tabulator selected the wrong analyte or test system and corrected it. The most frequent instance of this was when a tabulator selected the analyte "hemoglobin" instead of "Hgb, single analyte inst. w/self-cont." In most cases, however, the tabulator's entry was correct. In these instances, Westat created a file of records containing analyte and test system pairs not in the expanded Complexity Model and sent it to CDC for resolution. All new pairs involving existing analyte and test

system codes were subsequently added to the Complexity Model by CDC and so did not require further editing.

Data Validation Checks

Several validation checks were built into the Tabulation Device to ensure that data were collected. These included the following:

- No records could be entered with blank analyte names;
- No records could be entered with blank biological specimen names;
- No records could be entered with blank test system names;
- The number of quality control tests performed had to be less than the total number of tests performed; and
- The tabulation for a location could not be set to "Complete" unless a volume was specified for all listed clusters.

If a tabulator violated one of these constraints, the system displayed a diagnostic message and provided a prompt to correct the error.

Daily Log Sampling

Some sites provided test volume information in the form of daily logs. To avoid having tabulators enter over 300 sets of volumes for each cluster in such sites, the Tabulation Device had an automated sampling feature. To use it, the tabulator first indicated the periods during which a laboratory was open for business (e.g., January I to May 31 and September 1 to December 31), then entered which days of the week a laboratory was open for business and the number of annual holidays on which it was closed. The system then calculated the number of days during which the laboratory was open for business in 1996 and divided this number by 20 to obtain the sampling interval, *S*.

A day was randomly selected from among the first *S* days that the laboratory was open, and the first time period was set to start and end on that day. Every *S*th day thereafter that fell on a day of the week on which the laboratory was routinely open was defined as an additional time period until 20 evenly spaced days were selected.

These 20 one-day-long periods were the days for which the tabulator recorded data. If one of these days fell on a holiday when the laboratory was closed or for some reason had missing data, the tabulator could set a flag indicating that fact. The total number of days that the laboratory was open and the number of days for which data were collected were used at the home office when estimating the annual volume for each cluster.

Nursing Station Sampling

When listing the test record sites within a geographic location, the tabulator provided a unique name to identify each site; its address or other locating information, as appropriate; the name and phone number of a contact at the site; and the type of the site. Site types could be one of the following:

- Blood donor center;
- Bone marrow unit;
- Cancer center laboratory;
- Cardiology/open heart/cardiac cath;
- Clinic (e.g., urology, pediatrics, coagulation);
- Dermatology;
- Emergency room;
- Endoscopy service;
- Health screen/health fair;
- Helicopter;
- Hematology;
- Hemodialysis;
- Intermediate or short-term care;
- Interventional radiology;
- Labor and delivery;
- Main laboratory;

- Mobile unit;
- Neonatology;
- Nursing unit;
- OR laboratory;
- Other, specify;
- Outpatient delivery (e.g., home birth suites);
- Outpatient surgery;
- Paramedic; ambulance;
- Pharmacy;
- Physicians Office Laboratory (POL);
- Point of care;
- Recovery room (post-anesthesia care);
- Respiratory therapy;
- Rheumatology;
- Shock trauma;
- Sites where microscopes were used;
- Skilled nursing facility associated with hospital;
- Syndrome-specific (diabetes);
- Urgent care (nonemergency clinic);
- Urology;
- Visiting nurse;
- Women's clinic/OB clinic; and
- X-ray, nuclear medicine.

If "Nursing Station" was selected, a supplementary window appeared that let the tabulator indicate whether the nursing station was "Standard (Homogeneous)" or "Specialized." "Homogeneous" in this context

meant that the nursing station performed the same set of tests as one or more other nursing stations. Any other nursing station was, by definition, "Specialized." In hospitals with more than two homogeneous nursing stations, the system randomly selected two for data collection. The tabulator could enter data for the selected sites but not for the remaining homogeneous nursing stations. The total number of nursing stations and the number from which data were collected were used at the home office when calculating an estimate for the annual volume in nursing stations for each cluster. All "Specialized" nursing stations were tabulated.

3.1.1.4 Data Communications

Each laptop was equipped with a 33,600 baud Hayes modem for use in communicating with Westat's home office. The modem was used for sending and receiving e-mail text messages, for automatically sending data files to Westat, for updating the software, for sending new assignments to the field, and for sending assignment status data to Westat.

3.1.2 Data Processing and Data Quality Assurance

Data processing encompassed the tasks performed at the home office to convert the data received from the tabulators into a delivery file containing the list of analyte test system biological specimen clusters conducted under each CLIA ID number in calendar year 1996 and the number of each cluster that was performed for patient care. The following batch processing steps were used to process the data:

- Editing and cleaning;
- Annualizing daily logs;
- Summarizing site data to location level data;
- Summarizing location data to laboratory level data (CLIA ID number level); and
- Making special adjustments.

Each of these steps is discussed in a separate section.

3.1.2.1 Editing and Cleaning

The editing and cleaning process had several stages, including quality assurance checks; using machine edit programs to check for defined error conditions; printing and reviewing all records for each laboratory; and making any needed corrections.

Quality Assurance Checks

Sometimes analytes and test systems were discovered in the field that were not included in the expanded Complexity Model. When this occurred, the data tabulators were asked to use the comment field on the Tabulation Device to describe the new analyte or test system or, less frequently, the new specimen type. They also completed a paper form, the Tabulation Problem Sheet, which was included in the hard-copy case folder.[7]

The systems staff maintained a log of computer-related problems and reviewed the log to ensure that any needed data corrections were made. Westat systems staff also searched for two conditions that would indicate a computer problem:

- There should be no clusters with zero volumes for sites with a single time period and
- There should never be null volumes.

After these cleaning operations were complete, Westat converted the files to SAS datasets.

[7] A hard-copy case folder was provided to the tabulator for each location when it was assigned. It included laboratory contact information and administrative forms. See Chapter 4 for more details.

Using Machine Edit Programs to Check for Defined Error Conditions

- To ensure clean data, Westat created SAS programs to check that each dataset following conditions:
- There were no blanks in key fields;
- The Volume should be a positive number and greater than the QC Volume;
- The New Pair Flag should be " 1 " (true) if and only if the analyte code and test system code pair did not exist in the expanded Complexity Model; and
- Within a site, all time periods should have data with nonzero volumes.

While many of these checks duplicated those built into the Tabulation Device, examining them outside of the Tabulation Device using independently written code provided an extra quality check. Except for errors that were generated when blank records were included in the files, no record failed any of these tests.[8] The blank records were deleted from the Paradox files, which were then reimported and rechecked.

Printing and Reviewing all Records and Selected Sets of Records for Each Laboratory

- Westat printed the following set of reports for each location:
- All analyte and test system records;
- Records with Westat codes for test systems;
- Records with Westat codes for analytes;
- Records with blank analyte or test system codes; and
- Records with new pairs (other than those included in the above reports).

[8] Blank records were created by an error in a labor saving routine that automatically made analyte, test system, and biological specimen clusters available to a tabulator who had entered them for one time period when he or she was entering data for subsequent time periods. The routine always created records for 20 time periods (the maximum allowed) even if the tabulator specified a smaller number. The extra records contained blank analyte, test system, and specimen names, as well as blank volumes. This error was noticed and corrected early in the field period.

These reports were reviewed by medical technologists at Westat, leading to two types of changes. One change was to replace a specific Westat code (W-code) with the appropriate code from the Complexity Model wherever it occurred.

The other change was more local. When the medical technologists reviewing the data found that it was possible to replace a blank code with either a W-code or a Complexity Model code, the update needed to be made to a specific record, not globally. Similarly, in the few cases in which the reviewer found that the tabulator simply made a spelling error, it was necessary to correct a specific record.

3.1.2.2 Annualizing Daily Logs

Annualizing daily logs involved estimating the total annual volume for each record collected by subsampling daily logs. The resulting volume was called adjusted volume 1 (ADJVOL1) and was calculated by multiplying each volume entered by the inverse of the effective sampling fraction. For example, if a laboratory was open 315 days, and data were collected for 19 days, then the volume would be multiplied by 315/19 to obtain ADJVOL1. The adjustment factor for all sites that did not undergo daily log sampling was 1, so that the ADJVOL1 was the same as the volume for each cluster recorded at these sites. (See Section 5.1.2 for further discussion.)

3.1.2.3 Summarizing Site Data to Location Level Data

Sites were places within a laboratory location where a tabulator accessed records used during the survey. For example, in a hospital, records might be dispersed among the emergency room, pathology lab, a dialysis unit, and multiple nursing stations.

After daily log data were annualized, any identical clusters recorded within a site were combined by adding together their volumes so that each cluster was represented only once for the site. If nursing stations were subsampled at a given location, an adjustment was made for the sampling by multiplying each recorded volume in a homogeneous nursing station by the ratio of the total number of homogeneous nursing stations listed in the Tabulation Device to the number sampled. Thus, if a hospital had seven homogeneous nursing stations and three specialized nursing stations, data were collected from two of the homogeneous nursing stations and all three of the specialized ones. Adjusted volume 2 (ADJVOL2) was calculated by multiplying the ADJVOL1 in each of the

two tabulated homogeneous nursing stations by 7/2. In all other cases, ADJVOL2 equaled ADJVOL1 I.

To determine the total volume for each cluster at a location, the values of ADJVOL2 each distinct cluster were added together across all the sites within the location.

3.1.2.4 Summarizing Location Data to Laboratory Level Data (CLIA ID Number Level)

To determine the total volume for each cluster at a laboratory, the total volumes for each distinct cluster were added together, across all locations within the laboratory (i.e., locations within a CLIA ID number).

3.1.2.5 Making Special Adjustments

Because of special conditions associated with a few laboratories, some special adjustments were required (also see Section 5.1.2). In one large organization, 20 of 80 locations were randomly chosen and visited on site. The central organization reported the total number of analytes tested, but test systems varied from location to location. Because the sum of the volumes (by analyte) from the 20 laboratories was not equal to one-fourth of the total reported by the central laboratory, the volumes were adjusted in each of the 20 sampled sites by creating an adjustment factor equal to

$$\frac{\text{total volume reported by central laboratory for the analyte}}{\text{sum of the volumes from the 20 sampled sites for the analyte}}$$

The volume recorded for each cluster associated with the analyte was multiplied by the adjustment factor for that analyte. Processing then continued as normal for the data from each site.

In 10 other locations, from which Westat was able to obtain data for only a portion of the time that they operated during 1996, volumes were adjusted by multiplying the recorded volumes by the number of days that the location was open, then dividing by the number for which the tabulator collected data.

In three laboratories with many locations, a subset of the locations was sampled prior to assignment to the tabulators for data collection. The recorded volumes for these laboratories were multiplied by an adjustment factor equal to the total number of locations divided by the number of locations sampled.

3.2 Phase II-Collection and Processing of Data from Laboratories with Waived and PPM CLIA Certificates

Data collection in laboratories with waived and PPM CLIA certificates was performed by mailing a Test Inventory Form to the laboratories and then calling to collect the data recorded on the form. Medical technologists at the home office called the participating laboratories and collected the data by recording it on a NICLTS Telephone Data Tabulation Form. The forms were then keyed into ASCII files and independently key verified by entering the data a second time. The ASCII files were converted to SAS files, cleaned, and summarized.

This telephone data collection and subsequent key entry was analogous to the field data collection using the Tabulation Device in Phase I. Data processing of the keyed data was analogous to the processing of Phase I data after it was transmitted to the home office.

3.2.1 Data Collection and Entry

The NICLTS Telephone Data Tabulation Form contained one line for each waived and PPM analyte-test system pair in the Complexity Model. Complexity Model codes for the analytes and test systems were preprinted on the form. The form also had spaces for recording the biological specimen, biological specimen code, and calendar year 1996 volume. When the biological specimen was known in advance, the biological specimen name and code were also preprinted on the form. For example, the biological specimen for the Bionike A/Q Pregnancy Test for Urine was preprinted on the form as "Urine" with specimen code "30." For the remaining clusters, the tabulators entered the specimen name and code on the form. For all clusters reported by the laboratory, the tabulators entered the reported calendar year 1996 volume.

After the tabulator completed the data collection form, a supervisory medical technologist reviewed it for completeness, legibility, and accuracy of coding and counted the number of clusters on the form with volume data recorded. The supervisor recorded this count on the front cover of the data collection form next to the NICLTS ID for the laboratory.

Only numeric fields from the data collection forms were keyed. These fields were as follows:

- NICLTS ID - The 7-digit location ID (from the front cover);
- Count (from the front cover);
- Analyte code;
- Test system code;
- Biological specimen code; and
- Volume.

Each batch was independently key verified and copied to Westat's network.

3.2.2 Data Processing

Data processing of the Phase II data consisted of two major steps: editing and cleaning and summarization of multilocation data to laboratory level.

No special adjustments were needed for Phase IL

3.2.2.1 Editing and Cleaning

The data editing and cleaning process was similar to that performed for the Phase I data, but simpler. Because there were no reports of problems from the field that needed to be resolved, there was no need to locate

the files or records with known problems and correct them. Westat did, however, need to perform the remaining data editing and cleaning steps:

- Using machine edit programs to check for defined error conditions;
- Printing and reviewing all records and selected sets of records for each laboratory; and
- Making needed corrections.

Using Machine Edit Programs to Check for Defined Error Conditions

Westat wrote and applied a number of routine edits including the following:

- The value for the volume in each record in the mail/telephone data was greater than zero;
- The number of records entered for an ID matched the manual count recorded by the data collection supervisor on the cover of the data collection form;
- The set of NICLTS IDs included in a data file matched one-to-one with the set of IDs recorded in the Survey Management System as having completed tabulations; and
- All and only the NICLTS IDs sent to key entry were included in the data files.

Westat created a file of all clusters containing W-codes and another file of all clusters containing blank codes. A report was produced for Westat and CDC staff to use in assigning Complexity Model codes to these analytes and test systems. The report showed that over 25 percent of the blank codes were caused by the absence of one specific test system from the expanded Complexity Model. Senior medical technologists edited each of the W-codes and blank codes by checking all documentation in the relevant laboratory's permanent hard-copy file, reviewing the Complexity Model for more complete test system names, and double checking the hard-copy data tabulation form. When possible, the reviewing medical technologists used the more complete information obtained in this manner to correct the entries. The remaining records were sent to CDC for coding.

There were a number of reasons for the observed coding problems. Phase II information was gathered by telephone tabulators who could not directly observe test systems. Some respondents could not give complete information about test systems. Timing of the tabulation process may also have been a factor. Tabulation occurred approximately a year after the 1996 period for which test data were collected, and many of the test systems used

in 1996 were no longer in use. Respondents could recall use of a glucometer, for example, but not the name of the manufacturer. The small number of records with this type of partial information was also sent to CDC for coding.

3.2.2.2 Summarizing Multilocation Data to Laboratory Level

Summarizing of multilocation data to the laboratory level was considerably simpler for Phase II data than for Phase I data. Some laboratories had multiple locations from which Westat collected data using separate data collection forms. However, because staff at these locations compiled all the relevant data in one place, the locations did not need to be further subdivided into sites for data collection purposes. Furthermore, Westat did not perform either daily log subsampling or nursing station subsampling, and thus had no need to adjust the volumes to reflect such sampling.

Data summarizing to the laboratory level consisted of adding the volumes of like clusters from all the locations of a laboratory so that a single value was reported for each cluster found within a laboratory (i.e., within a CLIA ID number).