Electronic Data Processing, Analysis and Reporting for Public Health Surveys

Participant Manual

December, 2006

Acknowledgments

This manual was prepared by the United States Department of Health and Human Services Centers for Disease Control and Prevention (HHS-CDC), Global AIDS Program (GAP) Surveillance Team in collaboration with the World Health Organization (WHO), Geneva.

The original manual was written by Kimberly Marsh, MPH.

CDC/GAP thanks the following ministries of health for hosting pilot trainings:

<to be added>

Thanks also to UNAIDS and the surveillance and survey working group of the Office of the Global AIDS Coordinator (OGAC), consisting of:

- United States Census Bureau
- United States Agency for International Development (USAID)
- United States Department of Defense
- United States State Department

Table of Contents

Table of Contents, continued

Table of Contents, continued

Table of Contents, continued

Table of Contents, continued

Appendices

Notes

Introduction

Course Overview

What you should know before the course

This course is designed to provide basic technical skills in processing and analysing data, ultimately for the purpose of producing epidemiologic reports at the regional and national level.

To benefit from this course, you should be familiar with:

- the Microsoft Windows computing environment, (including moving, copying and renaming files and file folders)
- performing and interpreting both simple and more complex data analyses using either computer or paper-based statistical methods.

Familiarity with Epi Info is not required.

Finally, because antenatal clinic HIV sentinel surveillance is used as an example throughout this course, you should understand the basic approach to conducting such surveys in resource-limited settings. Become familiar with this type of surveillance before coming to class by reading the WHO Second Generation Surveillance Guidelines at (http://www.who.int/hiv/pub/surveillance/pub3/en/index.html) or other published literature.

Course purpose

The purpose of this course is to provide you with basic skills in data processing, analysis and report writing for survey data.

Specifically, the course will introduce best-practice techniques for systematically collecting, managing, processing and reporting HIV survey data from antenatal clinics (ANCs).

You will engage in the planning and implementation of the 2002 HIV sentinel surveillance round in a fictitious country called Suri in order to understand and apply these best-practice techniques.

Course objectives

By the end of the course, you should be able to:

- design easy-to-use data collection and electronic data-entry forms
- develop simple and complex check code to validate data entry
- oversee and perform data entry
- develop and document data cleaning and database storage strategies
- conduct simple exploratory analysis for data cleaning purposes
- clean and prepare data for analysis
- Perform simple and complex descriptive analyses
- develop clear and concise national and regional reports.

Operation System and Epi Info Software Requirements

Epi Info [for Windows] is a public domain software package designed for the global community of public health practitioners and researchers. It provides for easy form and database construction, data entry and analysis with epidemiologic statistics, maps and graphs. Epi Info should be pre-loaded on classroom computers and can be accessed by double-clicking the icon on the computer desktop screen.

On your desktop, double-click the icon: Epi Info	**Note:** If Epi Info is not loaded onto your computer, you can either request a copy by CD-ROM or download the latest version from http://www.cdc.gov/epiinfo/downloads.htm. Directions for installing the software are also available from this site.

System requirements for Epi Info

- Windows 98, NT 4.0, 2000 or XP
- A minimum of 32 MB of Random Access Memory is recommended for Windows 98, 64 MB minimum for Windows NT 4.0 and 2000 and 128 MB minimum for Windows XP
- A 200 megahertz processor (recommended)
- At least 260 MB of free hard disk space (Drive C) to install; 130 MB after installation

Training Schedule

The course lasts five days. We plan to cover all exercises, Exercises 1 through 12, during the duration of the week. Additional group activities, such as the development of a data screen for a country-specific ANC form, sample national reports or a PowerPoint presentation (as described in Exercise 1 and Exercise 12) may require additional time, and may be condensed in the interest of time. See your course materials for a copy of the course-specific training schedule.

Course Schedule

Day 1	Day 2	Day 3	Day 4	Day 5
☑ Course Overview ☑ Exercise 1 ☑ Exercise 2	☑ Exercise 3 ☑ Exercise 4 ☑ Exercise 5	☑ Exercise 6 ☑ Exercise 7	☑ Exercise 8 ☑ Exercise 9	☑ Exercise 10 ☑ Exercise 11 ☑ Exercise 12 ☑ Final Test ☑ Course Evaluation

Using the Hints and Directions

Watch for the icons below. They will assist you by pointing out hints or directions.

1. A note icon is used to draw your attention to key information (🗒)

   Example: 🗒 Note that you may need to…

2. A light bulb icon marks key information to aid in understanding how Epi Info works (💡).

   Example: 💡 Epi Info can also…

3. Activities for practising the skills you've learned are characterised by the heading,

   Try it yourself!

Using the Hints and Directions, continued

4. Command buttons, check boxes and radio buttons are capitalised and bold.

   Example: Click **Cancel**.

5. Dialog boxes and other windows requiring user interaction are capitalised with a bold text.

   Example: A **Field Definition** dialog box appears.

Additions, Corrections, Suggestions

Do you have changes to suggest for this module? Is there other information you'd like to see? Please email Alison Smith, the instructor.

We will collect your emails and consider your comments in the next update to this module.

Email address:

cfq2@cdc.gov

Exercise 1

Designing Easy-to-Use Forms

Overview

What this exercise is about

You have been asked to assist the HIV sentinel surveillance team in documenting and improving the existing Epi Info 2001 Antenatal Clinic sentinel surveillance system in Suri. The goals of the review are to ensure appropriate data collection, entry, analysis and reporting for the upcoming 2002 round of sentinel surveillance.

What you will learn

By the end of this unit, you should be able to:

- identify the steps involved in designing good data collection forms
- apply knowledge of good design techniques to design a sample data collection form.

Resources

Appendix A – Country-specific HIV Surveillance Data Collection Forms
Appendix B – HIV Surveillance Data Collection Form for ANCs - WHO recommended
Appendix C – Suri Surveillance Data Collection Form for ANCs (YR.2001)
Appendix D – Suri Surveillance Data Collection Form for ANCs (YR.2002)

Designing Forms

Good form design is critical to ensuring that data collected during the survey accurately reflect the responses provided by the patient or the medical staff. Here are the steps we will follow when designing survey data collection forms.

Form design steps

You will have a chance to do each of these activities:

1. Review previous survey data collection forms or forms used previously in your country or in other countries.
2. Generate a rough-draft list of all variables that you want to include in your survey and their possible responses.
3. Create a flowchart of variables, eliminating redundant variables or adding variables or directions for clarification.
4. Group and order variables depending on when and by whom they are collected.
5. Develop a rough draft of the form using best-practice design principles.

Follow the steps in Exercise 1 to better understand the principle methods, tools and techniques for designing easy-to-use forms. At the end of the exercise, compare your form to the WHO Recommended Ministry of Health HIV Surveillance Data Collection Form for ANC Clinics in Appendix B and Suri's 2001 Form.

Case Study: HIV Sentinel Sites, Suri, 2002

Please read the Suri case study in preparation for discussion afterwards.

Suri case study

Suri is a fictitious country that, as recently as 1999, had very limited data about the prevalence of HIV in the country. A survey among commercial sex workers (CSWs) conducted by a local non-governmental organisation (NGO) in four of the five regions in Suri in 1998 showed HIV prevalence ranging from 35% in Tibul to 48% in Ashra. A convenience sample of tuberculosis patients obtaining directly observed therapy in 18 clinics in those same regions demonstrated high co-morbidity between TB and HIV in 1999. Of the 765 patients infected with TB, 596 (78%) also tested positive for HIV.

Suri case study, continued

Based on the results of the 1998-1999 surveys in the special population groups, the Minister of Health (MoH) in Suri tasked the national HIV/AIDS surveillance team to establish an HIV sentinel surveillance system among pregnant women to further describe the HIV epidemic in the country. Antenatal-care sentinel surveillance is one of the primary tools in a generalised epidemic for estimating HIV prevalence among pregnant women. Results from the survey can aid in the description of the number and demographic characteristics of HIV-infected pregnant women at their first attendance at participating clinic sites during the survey period. Survey data can also be used to longitudinally monitor trends and changes in infections, as well as to assess the potential impact of targeted programmes and interventions among these women. In some instances, these data can be used to estimate HIV prevalence among the general population and project infection levels in the country over the next 5-8 years.

In 2000, HIV sentinel surveillance data were collected in 19 sites in four of the five regions in Suri. Data were collected on hardcopy forms and then entered using the software tool Epi Info 6. Paper copies of the 2000 data collection forms are no longer available and three sites never submitted their results; however, a data file of the line-listed records for those sites that submitted is still accessible electronically. Results from the 2000 ANC round were never disseminated in a national report, although the Minister of Health reported that 32.4% of pregnant women aged 12-49 and sampled during the ANC survey were HIV-infected. This figure established Suri as having one of the highest HIV burdens in the world. To prevent further spread of the epidemic, the MoH, in collaboration with local NGOs, launched a major Information and Education Campaign to combat the high infection levels.

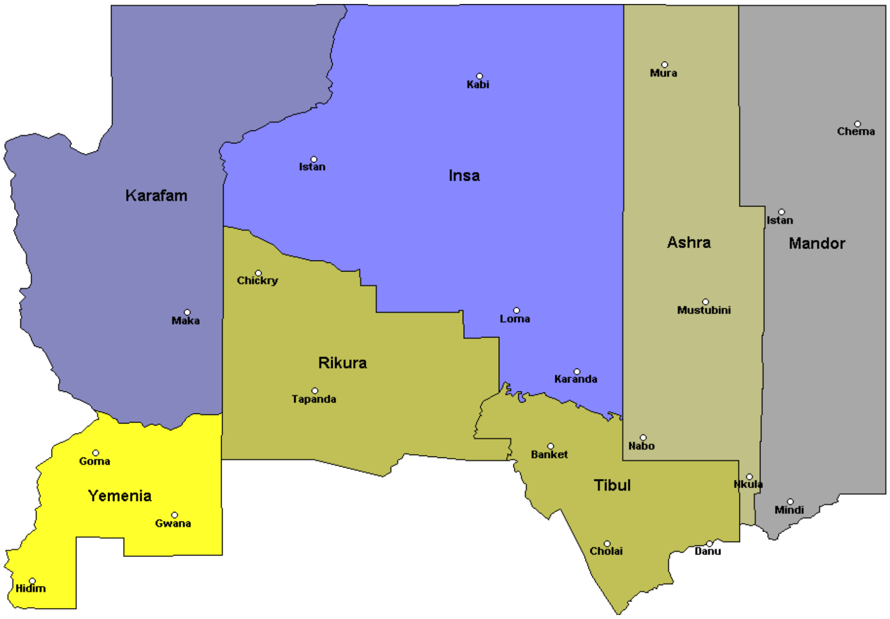
In 2001, the HIV/AIDS surveillance team repeated the survey at the request of the Minister of Health in preparation for an upcoming presidential address to the Africa region on the AIDS crisis. During the second round, the HIV surveillance team expanded the number of sites to include three additional clinics. Once data were collected, the Suri MoH Statistics Team created an Epi Info HIV sentinel surveillance information system for entering and analysing the HIV survey data. Again, no national report was produced; however, the government announced a decline in HIV prevalence from 32.4% in 2000 to 31.8% in 2001 among pregnant women. Although recognising that HIV prevalence was still high in Suri, the president highlighted the effective response that the government was making to control the epidemic.

Suri case study, continued

Since the conclusion of the 2001 survey, the MoH has been eager to further assess the impact of ongoing prevention efforts in the country. For the 2002 round of ANC surveillance, the minister hired a team of consultants to assist the HIV Surveillance team in more rapidly collecting, managing and analysing the 2002 HIV sentinel surveillance data. In addition, they have asked the consultants to oversee the design and dissemination of the first national report describing the 2002 sentinel surveillance results and the HIV prevalence trends from 2000-2002 in Suri. The consultants accepted the task of working with the HIV Surveillance Team, with the condition that they be able to review the previous data collection, management and analysis procedures and to suggest areas for improvement in the upcoming 2002 round.

In the exercises that follow, you, as a new epidemiologist in the HIV Surveillance team, will join the consultants in Suri (i.e., your instructors) as they plan for the upcoming round, process and analyse the results, and create a national report for dissemination. The exercises will lead you through a process of critiquing activities in 2000 and 2001 and planning for activities in 2002. Shortly after you complete the planning process, data for the 2002 round will be gathered according to the team's recommendations. You will then assist the consulting team in preparing a file for data analysis. In the final exercises, you will analyse the data for the year 2002 and work with colleagues to produce the first national HIV sentinel surveillance report that summarises the state of the HIV epidemic among pregnant women from 2000 to 2002 in Suri.

Map of HIV sentinel sites, Suri, 2002.



Form Design Steps 1 and 2

Step 1: Review data collection forms

Let's look at the first two steps of the form design process now.

1. Review previous survey data collection forms or forms used previously in your or in other countries.

It is useful to identify all existing forms that are in use or have been used in your country or other countries. Often, reviewing previous data collection forms with others or discussing the variables of interest can give you a better understanding of what to do and what not to do in order to facilitate data collection.

In addition, a review of previously developed forms or forms used elsewhere can give you a better understanding of the data that might be useful. It's important to talk with people who have collected administrative, demographic and laboratory variables different from what you collect. A variable that might work in theory may be difficult in practice to collect or use.

Step 2: Generate a rough-draft list of variables

2. Generate a rough draft list of all variables that you want to include in your survey and their possible responses.

You should identify all of the variables that you may want to collect on the form and their possible responses. Do not forget to include variables on the form that are administrative in nature, such as clinic location or form identifier variables.

Consider how you will ask the question; for example,

- Will you ask for the mother's age in years or for her date of birth?
- Will you ask for an overall positive or negative HIV status, or will you ask for each of the test results that can be used to determine a positive or negative diagnosis?

It is important to consider this in advance to determine what additional analyses may have to be done during the post-data collection period.

Step 2: Generate a rough draft list of variables, continued

In addition to noting all possible variables and responses, you should identify the ways that you might validate the response for each variable; for example, for date of birth, you might limit the year variable during data entry to only those years during which an eligible mother could be born. You might also want to specify which variables are required and to consider how missing or unknown values will be indicated on the form.

Try it yourself!

Activity 1, Review Survey Forms and Generate List of Variables

a. Do you have questions on the case study?

b. Look at Appendix A. You will need this information for the following exercise.

c. Create a sample ANC data collection form for Suri using the following steps:

- Refer back to Steps 1 and 2.
- List all the variables you want to include.
- Define response values as outlined in Step 2 that are appropriate to the variable being considered (e.g., a list of occupations for the occupation variable).

Form Design Steps 3 and 4

Step 3: create a flow chart of variables

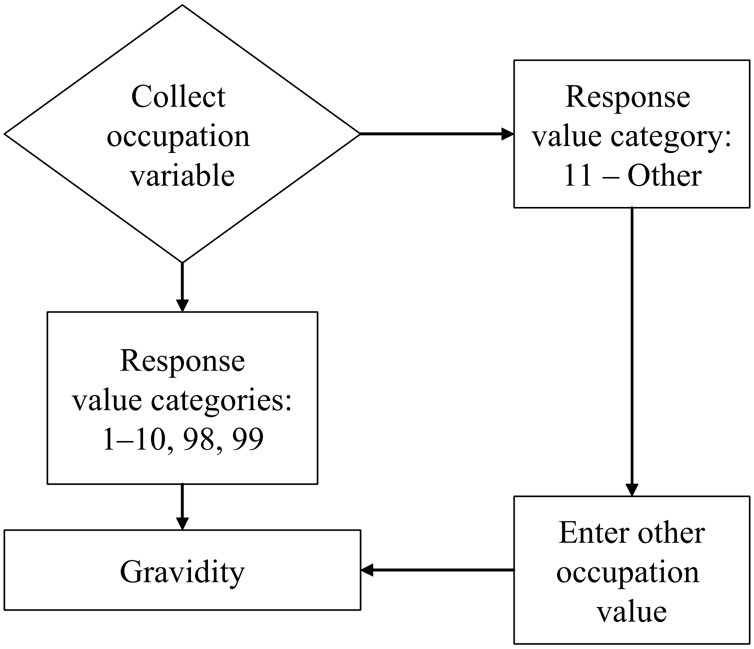
3. Create a flow chart of variables, eliminating redundant variables or adding variables or directions for clarification.

Review the variables to determine if any of the responses to these variables depend upon or affect answers to other variables in the form. These types of linked questions are also known as navigation variables.

Skip variables

One example of a navigation variable is a skip variable. An example includes asking the user to write in a woman's occupation when the response "11 – other" is checked in the occupation field. If the woman's occupation is not "11 – other," this variable can be skipped and the collector or data-entry person can enter the next value for gravidity. The flowchart might look like Figure 1.1 below.

Figure 1.1. Using skip variables.


Collect occupation variable
Response value category: 11 – Other
Response value categories: 1–10, 98, 99
Gravidity
Enter other occupation value

Cascade variables

Another type of navigation variable is a cascade variable. A cascade variable may limit the collection of unnecessary data, since once you know the answer to that variable, other variables can be derived. For example, in your database, you should already have clinic locations linked to districts and provinces; therefore, if you know the clinic location, the data for districts and provinces do not need to be collected.

Step 4:
Group and order variables

4. Group and order variables depending on when and by whom they are collected.

Grouping variables according to the person with responsibility for collecting the data or how data naturally arise in the course of care will ensure that variables are collected more accurately. For example, placing the demographic variable groups after the laboratory test results on the form may not be appropriate if laboratory testing is being conducted centrally and demographic data collection occurs first. You should consider the order and grouping of variables according to tasks and when and by whom they are collected when placing variables on a form.

Mapping out decisions about when, where and by whom data are collected in your flow chart should be noted. During the design of the data collection form, you may also wish to note who has responsibility for collecting specific variables and where they will be collected in the flow chart you just created. For example, a nurse may collect the demographic data, but the laboratorians may receive the form to complete the test results. On your form, it may be helpful to include instructions directing the nurse to ensure that all variables in the demographic section are completed prior to sending the form to the laboratory.

Try it yourself!

Activity 2, Create a Flow Chart of Variables

a. Create a flow chart of variables, eliminating redundant variables or adding variables or directions for clarification, as described in Step 3 above.

b. Group together variables based on where, when and by whom they are likely to be collected as outlined in Step 4. To do this, you may need to make some assumptions about the type of staff and locations that are available in the areas in which the survey will be conducted.

Form Design Step 5

Step 5: Develop a rough draft of the form

5. Develop a rough draft of the form using best-practice design principles.

Data collection forms should be designed with the data collection and entry staff in mind. To best meet their needs, the following form design principles should be considered:

Form design principles

A. Display only the minimum instructions and data labels on forms.

As a supplement to the data collection form, create additional training materials that clarify the data collection form variables rather than including additional text instructions on the form.

The problem: A poorly formed variable question takes up space.

Mother's age at the time of first visit in years: _____

The solution: With appropriate training material specifying this variable as the age of the mother, listing a variable labeled ‘Age in years’ will be clearer.

Age: _____ years

Form design principles, continued

B. Use as much 'white space' as possible.

Crowding variables and their responses together, or limiting the area in which a text response can be written, may make it difficult to read the data correctly.

The problem: Poor use of white space makes it difficult to write text.

Form ID: ____________ Clinic site: ___________ District:___________

Province: ____________

The solution: Identify the coding scheme for the form ID and pre-print this in the corner. Collect clinic site only, since District and Province can be obtained from the form. If the form ID includes the clinic location, this variable can also be eliminated.

C. Clearly and consistently locate variable labels and their responses.

Variable labels should either precede, or be followed closely by, their responses. If a text response is expected, a long line with enough white space to write the response should be used after the variable label. Further, throughout the form, there should be consistent sequencing of a label and then a response, except for checkboxes and radio buttons which should always precede the label.

The problem: A confusing checkbox will lead to inaccurate selections.

Age: 15-19 □ 20-24 □ 25-29 □ 30-34 ☑ 35-39 □ 40-44 □ 45-49 □

Boxes in the middle may be mistakenly checked.

The solution: Stack responses rather than list them across the page.

AGE:
□ 15-19
□ 20-24
□ 25-29
☑ 30-34
□ 35-39
□ 40-44
□ 45-49

Form Design Step 6

Step 6: Conduct usability testing

6. Conduct usability testing with personnel responsible for data collection.

Prior to distribution, test your draft form in a sample of sites by different personnel who have responsibility for data collection during the survey. Iterative form design based on user feedback is the most critical issue in ensuring that unexpected and correctible errors are not introduced into the data during the collection period.

Designing easy-to-use data collection forms is the first step in ensuring the accuracy of data collected during the survey. The role of the form should be to guide data collectors as they fill it out while reducing or eliminating errors and inappropriate responses.

Try it yourself!

Activity 3, Develop a Rough Draft Form

Develop a rough draft form as discussed in Step 5.

Activity 4, Compare Your Form with the WHO Recommended Form

Rather than pilot-test the draft form, compare your form to the WHO Recommended Ministry of Health HIV Surveillance Data Collection Form for ANC Clinics in Appendix B and Suri's 2001 Form in Appendix C. Note the similarities and differences. If you have access to your country's ANC form, compare this form as well.

The Ministry of Health sample data collection form includes the recommended variables and responses. Individual countries, as you have seen from your review of the ANC forms, may choose to adapt this form to local needs.

Activity 4, Compare Your Form with the WHO Recommended Form, continued

After discussion with the consultants, the surveillance team has decided to modify the 2001 form to collect additional data that may be useful during analysis for 2002. The final 2002 HIV Surveillance Data Collection Form for ANC Clinics is shown in Appendix D and can be compared with the 2001 HIV Surveillance Data Collection Form for ANC Clinics shown in Appendix C. Which three variables will be added to the 2002 form?

a. ____

b. ____

c. ____

Activity 5, Redesign a Form

Choose one form from the samples provided and redesign it based on the design best-practices discussed in this exercise. Look for:

a. redundant data
b. unclear format
c. any other ways to improve.

Exercise 2
Designing Data-Entry Forms

Overview

What this exercise is about

Suri's MoH Statistics Team created an electronic ANC data-entry form for the 2001 ANC survey using Epi Info. For the upcoming 2002 round, the HIV Surveillance Team has decided to expand data collection to include additional syphilis testing variables that are noted in Exercise 1. As a result, the 2001 data-entry screen must be modified to add three variables:

- RPR Test Date
- TPHA Syphilis Result
- TPHA Test Date

Follow the steps in Exercise 2 to assist the consultants in modifying the existing 2001 system and in documenting the changes in preparation for the 2002 survey round. Once modified, the 2002 system will be used centrally by the MoH to enter data.

What you will learn

At the end of this exercise, you will be able to:

- define and understand the relationship among projects, views and tables in Epi Info
- construct a data dictionary that documents the types of variables in the electronic database
- explain the difference among data types (e.g., text, numbers and dates) and how they are used in Epi Info
- document variable entities, attribute names, variable prompts, descriptions, values, types and character lengths in a data dictionary
- add variables and legal values to the questionnaire
- save the view.

Starting location

Epi Info Main Menu

Overview, continued

Resources

Appendix D – Suri Surveillance Data Collection Form for ANC (YR.2002)
Appendix E – Data Dictionary for the Suri 2001 ANC Survey

Overview of Epi Info Make View

Epi Info Project

Data-entry screens are the visual interface between a computer user and the database where data are stored. Epi Info uses the Microsoft Access file format. The file, called a **Project**, organises information contained in a system, including:

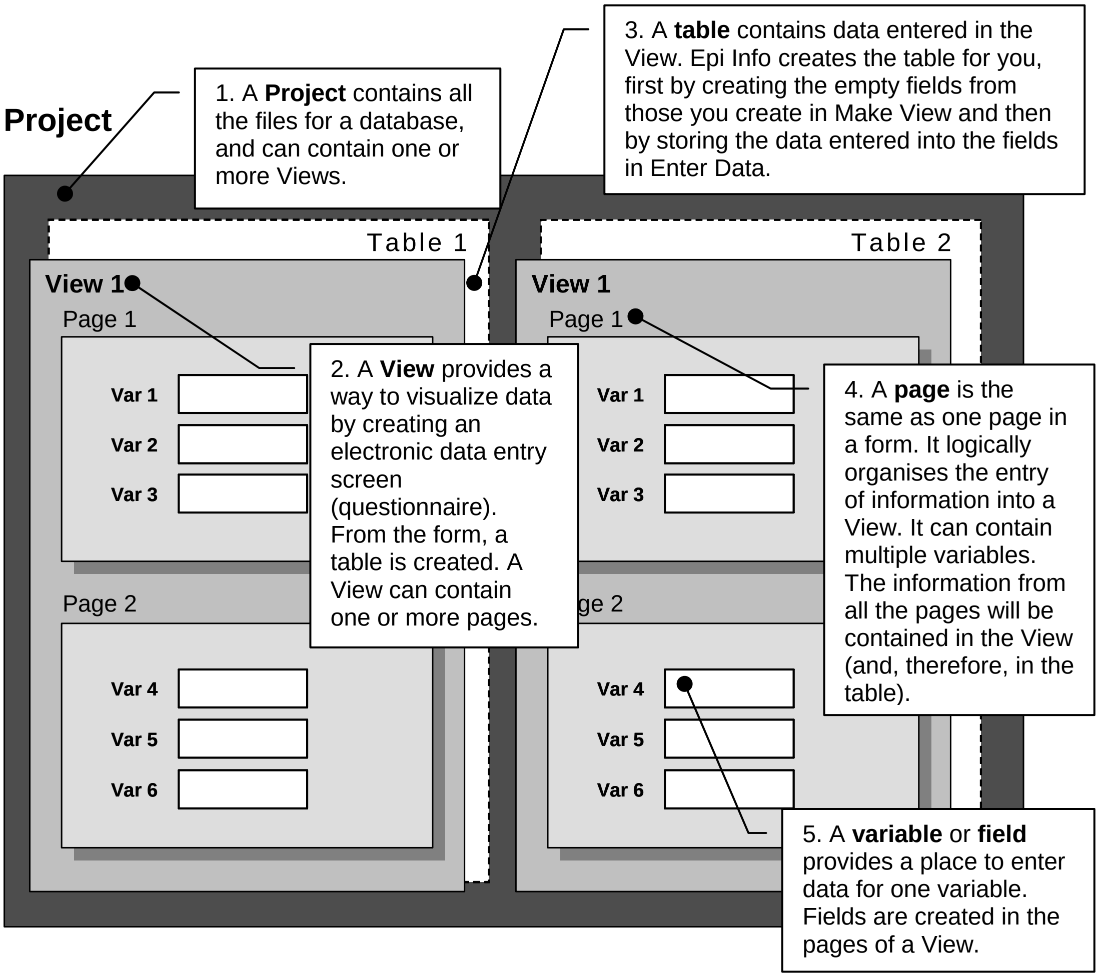
- the data-entry screen(s)
- rules for entering data
- the database proper.

For the Epi Info ANC surveillance system in 2001, the project was called ANC2001 using Epi Info's Make View application.

Relationship among projects, views, tables and variables

In Epi Info Make View, a **Project** contains one or many data-entry screens (e.g., for entering sentinel surveillance data), which are also called views. Each **View** contains information about one data table. **Data tables** often include information about variables to be collected in the View. The following diagram may be useful in showing the relationship between the Project, View, table and variable:

Relationship between Projects, Views, tables, and variables, continued



1. A **Project** contains all the files for a database, and can contain one or more Views.

2. A **View** provides a way to visualize data by creating an electronic data entry screen (questionnaire). From the form, a table is created. A View can contain one or more pages.

3. A **table** contains data entered in the View. Epi Info creates the table for you, first by creating the empty fields from those you create in Make View and then by storing the data entered into the fields in Enter Data.

4. A **page** is the same as one page in a form. It logically organises the entry of information into a View. It can contain multiple variables. The information from all the pages will be contained in the View (and, therefore, in the table).

5. A **variable** or **field** provides a place to enter data for one variable. Fields are created in the pages of a View.

Creating or modifying views

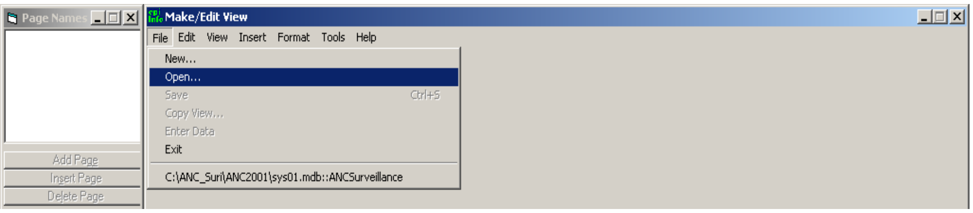
In Epi Info, new Views (i.e., data-entry screens) can be created in the application tool Make View. Existing views can also be modified with this tool. Make View is accessible in Epi Info either through the main menu or as a button on the start-up screen.

Viewing ANC 2001

To view the 2001 ANC sentinel surveillance Epi Info electronic data-entry form in Make View:

1. From the Epi Info main menu, click the **Make View** button. A window with a menu bar and blank area for creating a view is displayed.

2. Click on **File** in the menu bar and then click **Open...**.



3. In the Select the Project dialog box, type *C:\ANC_Suri\ANC2001* or use the drop-down box next to Look In to:

   - Select the C:\ drive
   - Double-click the ANC Suri folder name to open
   - Double-click the ANC2001 folder name to open.

4. Click on the file name *sys01.mdb* to select it.

5. Click **Open...**.

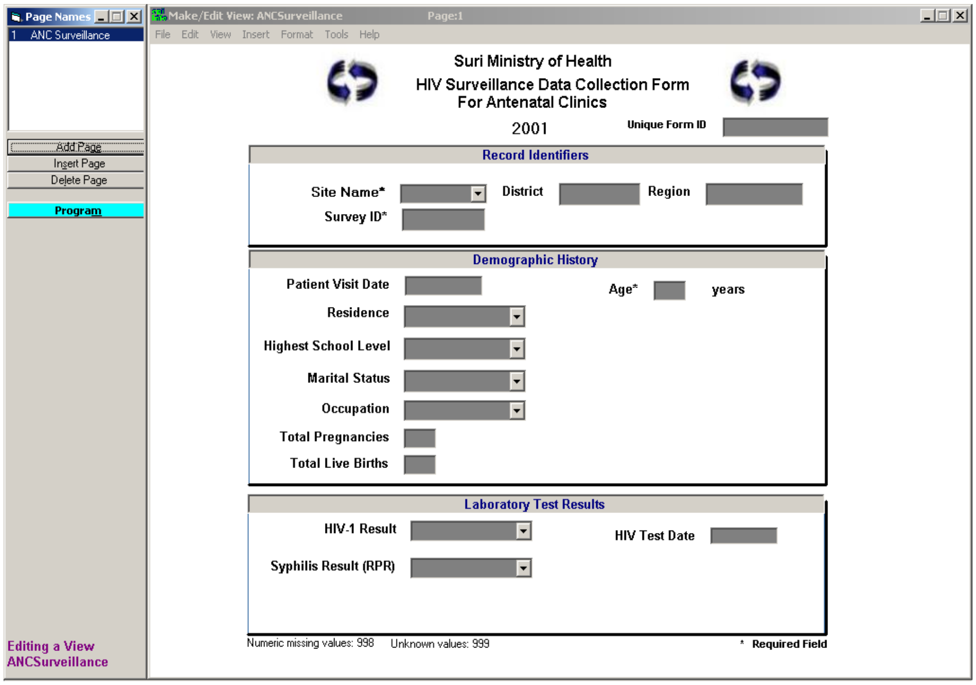
6. In the **Select a View** dialog box, click the view, *ANCSurveillance*.

7. Click **OK**. You have now opened the 2001 data-entry screen ANCSurveillance in the Project sys01.mdb.

Make View's file menu remembers the last view that was opened. This file can be easily accessed again by clicking on the File menu and selecting the file name at the bottom of the menu rather than opening the project and view again.

Surveillance system in Epi Info

The Epi Info 2001 ANC sentinel surveillance system should appear as pictured below:



Suri Ministry of Health
HIV Surveillance Data Collection Form
For Antenatal Clinics
2001

Unique Form ID

Record Identifiers

Site Name* | District | Region
Survey ID*

Demographic History

Patient Visit Date | Age* years
Residence
Highest School Level
Marital Status
Occupation
Total Pregnancies
Total Live Births

Laboratory Test Results

HIV-1 Result | HIV Test Date
Syphilis Result (RPR)

Numeric missing values: 998 Unknown values: 999 * Required Field

Capturing data

The tools provided through Make View allow you to develop and/or modify the View to electronically capture data during entry.

Documenting Your Data-Entry Form Using a Data Dictionary

Data dictionaries

Data dictionaries are critical in transitioning from a paper-based form, like we created in Exercise 1, to the actual electronic data-entry screen or View. Systems developers often document aspects of a system, and the data-entry screens in particular, using a *data dictionary*.

Data dictionaries, continued

A data dictionary is an electronic file that describes the basic organisation of a project or database. Data dictionaries can be part of the electronic database or they can be described separately in a word processing document. Data dictionaries should contain all of the rules that guide data entry and should be available to all users as part of the system documentation.

The terms "field" and "variable" are used interchangeably throughout this course and in Epi Info. They refer to the prompt or the data-entry box, or the name used to reference the data stored as the response to a question.

Components of a data dictionary

In a data dictionary, the following descriptive information is typically included:

- **Entity.** Describes a superset of the variables (such as location, identifiers, demographic or laboratory variables) that serve a similar purpose or are alike.

- **Variable Prompt.** The prompt/question label on the screen (if any) and a layman's definition of its meaning if that meaning cannot be understood from the field name. For example, the variable prompt of age is the age (years) or age of the mother in years at the time of first visit.

- **Type.** Variable type refers to text, date, numeric or other variable descriptors. If a field is numeric, only numbers are allowed. If the field is character-based, any characters (including numbers) are permitted to be entered. However, the variable during analysis will be treated as a character value, regardless of the input.

- **Size.** Size describes the number of characters or values that can be entered as a value. For example, age can be considered a numeric-type variable with a length of 3.

- **Field Name.** The name of the variable/field in the dataset where data, such as age, are to be entered.

Components of a data dictionary, continued

- **Code Table Values.** Code Table Values are those text or numbers that are acceptable in the response. They may include either a range of numbers (if the variable is numeric) or a specific set of text responses (if the variable is text.) For example, values for age may be 15-49 or 998–missing or 999–unknown. Values for region include: “MVG”– Mavinga, “MAS”–Masana, “HAR”–Hatar, and “MAN”– Malange.

- **Comments.** Comments may refer the user to other documentation (such as check code) or may note a discussion that led to recommendation of the variable or of specific values. In addition, they can indicate when missing or unknown values should be used or how values, when skipped, will be stored in the database.

- **Version Control.** Version control lists the date when the variable was first collected and the date when it was no longer collected, where applicable.

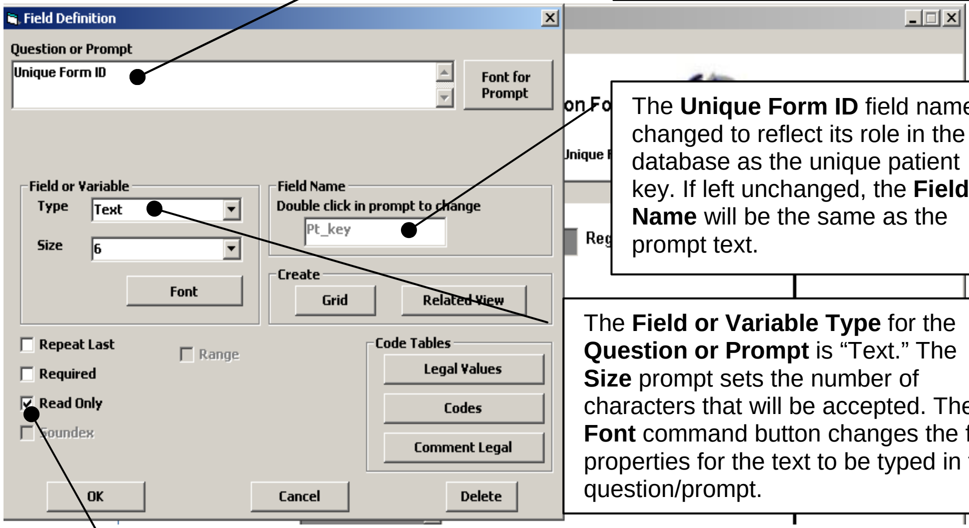
Other descriptive information may be documented as well, depending on the software used to create the data dictionary. A data dictionary created for the Suri ANC sentinel surveillance system is accessible in Appendix E.

Confirming the Data Contained in the Dictionary with that on the Screen

1. Double-click the label or prompt **Unique Form ID** on the *ANCSurveillance* view.

2. A **Field Definition** dialog box appears.

Field definition dialog box



Question or Prompt: Unique Form ID

Field or Variable
Type: Text
Size: 6

Field Name
Double click in prompt to change: Pt_key

☐ Repeat Last
☐ Required
☑ Read Only
☐ Soundex
☐ Range

The **Question or Prompt** for the Unique Form ID variable is “Unique Form ID.” The **Font for Prompt** command button shows the font, size and style of the prompt.

The **Unique Form ID** field name is changed to reflect its role in the database as the unique patient key. If left unchanged, the **Field Name** will be the same as the prompt text.

The **Field or Variable Type** for the **Question or Prompt** is “Text.” The **Size** prompt sets the number of characters that will be accepted. The **Font** command button changes the font properties for the text to be typed in the question/prompt.

The **Read Only** checkbox indicates that the data for this field cannot be entered or modified during data entry. The values will be assigned based on values from other fields on the View.

Numeric field or variable types

*Choosing the Field or Variable type **Label** in the Field Definition box will create a prompt on the data-entry screen without a data-entry box.*

In the Field Definition box, the Field or Variable Type Numeric has a default pattern of two numbers, represented by two “hash” signs (##). Any valid numeric format is acceptable. For numeric fields requiring decimal points, the hash signs can be entered before and after the period (e.g., ##.### for up to three decimal places.)

For Text types using drop-down boxes, allowable values can be seen by clicking on the down arrow in the box next to the prompt on the main Make View screen or by clicking on the non-greyed Code Table command button in the Field Definition Box.

Try it yourself!

Activity 1, Review the Suri 2001 Variables

Review the variables in the 2001 electronic system and compare them to the Suri data dictionary located in Appendix E, following Step 1 above. Fill in the spaces for cells containing question marks using the data-entry form and the information contained in the field definition window of that variable on the screen.

In the Comments section of the data dictionary, an explanation of how to interpret and add to these will be provided in later sections and exercises; therefore, you can leave this column blank for now.

Creating a New Project and View

You will need to save the View as a new file to avoid overwriting the 2001 system and data. To do this, follow the steps below:

Saving changes to the view

3. From the File menu, select Copy View.

Epi Info automatically saves changes to the View; however it is good practice to save your changes periodically and before exiting as well.

4. Make sure that the default value (e.g., the third option) is selected to **Make new View only.** A new empty data table will be made automatically during first data entry.

5. Make sure that the **'Copy Code tables or links in same MDB'** is checked. This will ensure that all of the values for the drop-down boxes are included in the new system. Click **OK**.

6. Click on the ANC2002 folder if it is not already highlighted. Type *sys02.mdb* into the File Name prompt.

7. Click **Open...**. You will be prompted to name the view.

8. Click **Yes** to acknowledge that copying will overwrite all code tables and relates.

9. Click **OK** on the message stating that the copy was successful.

Saving changes to the view, continued

You have now developed and saved the 2002 ANC sentinel surveillance system data entry screen system *(sys02.mdb)* file with the *ANCSurveillance* view in the *C:\ANC_Suri\ANC2002 folder*.

Opening and making changes to a project or view

1. From the **File** menu, select **Open....**

2. On the dialog box that opens, click **Change Project**.

3. Navigate to the *C:\ANC_Suri\ANC2002* folder.

4. Highlight the *sys02.mdb* project and click **Open...**. Select the *ANCSurveillance* view. Click **OK**.

Adding Variables to the Questionnaire

Additional variables will need to be added or modified in the system based on the changing needs of public health. In Suri, the consultancy team would like to take advantage of increased accessibility to TPHA syphilis testing during the 2002 survey. These variables have been added to the 2002 form already. It can be viewed in Appendix D.

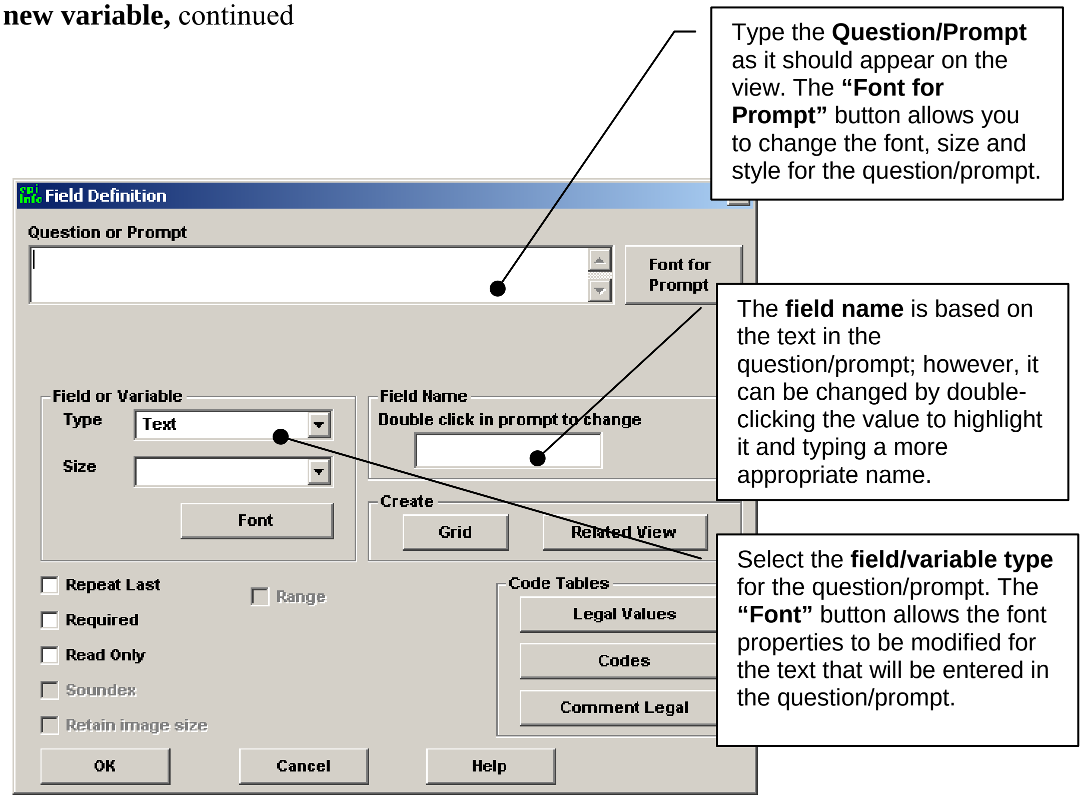
Steps to add a new variable

To add the TPHA variable:

1. Right-click on the view where you want to add the TPHA result variable. Use the data-entry form in Appendix D as a guide to its placement on the view.

A blank **Field Definition** box will open.

Steps to add a new variable, continued



Type the **Question/Prompt** as it should appear on the view. The **"Font for Prompt"** button allows you to change the font, size and style for the question/prompt.

The **field name** is based on the text in the question/prompt; however, it can be changed by double-clicking the value to highlight it and typing a more appropriate name.

Select the **field/variable type** for the question/prompt. The **"Font"** button allows the font properties to be modified for the text that will be entered in the question/prompt.

2. Use the following information to add the new syphilis result variable:

   a. Type into Variable **Question/Prompt**: *Syphilis Result (TPHA).*
   b. Click on the **Font for Prompt** button to change the font. Select: Arial, Bold, Size 10.
   c. In the **Field or Variable** type, select Text, Size: 2.
   d. In the **Field Name** box, highlight and delete the existing field name. Type in the new field name: *TPHA_res*. This step must be done when the field name box is first open.

3. Select **OK** to exit the Field Definition box.

Try it yourself!

Activity 2, Place Additional Variables in the Form

Repeat Steps 1 through 3 to add these additional variables to the form.

Question/ Prompt	Field Type	Pattern	Font	Field Name
RPR Test Date	Date	DD-MM-YYYY	Bold, Arial, Size 10	Rpr_date
TPHA Test Date	Date	DD-MM-YYYY	Bold, Arial, Size 10	Tpha_date

Also, change the System Heading label to show that this is the 2002 system rather than the 2001 system.

Creating Legal Values for Variables

You may have noticed when creating the *TPHA_res* variable that the text field prompt that was created allowed the data-entry person to type any text into the box. For example, a positive result for syphilis in the TPHA test could be entered as:

- “1 – Positive”
- “1-Pos”
- “Positive”
- “1”
- “Yes”
- “Don't know”

or any other text response, depending on the person typing in the data.

Epi Info code tables

To eliminate the potential errors introduced by allowing free text entry and to speed data entry by giving data-entry staff a choice, we use **Code Tables** in Epi Info. As you saw previously, variables for entering occupation, residence, marital status and HIV result all reference a specific Code Table containing allowable values.

Epi Info code tables, continued

Code Tables can be created in the **Field Definitions** box by right-clicking on the variable for which you want to create the drop-down lists. To create a Code Table for the syphilis variable (*TPHA_res*) containing three acceptable values, follow the steps below:

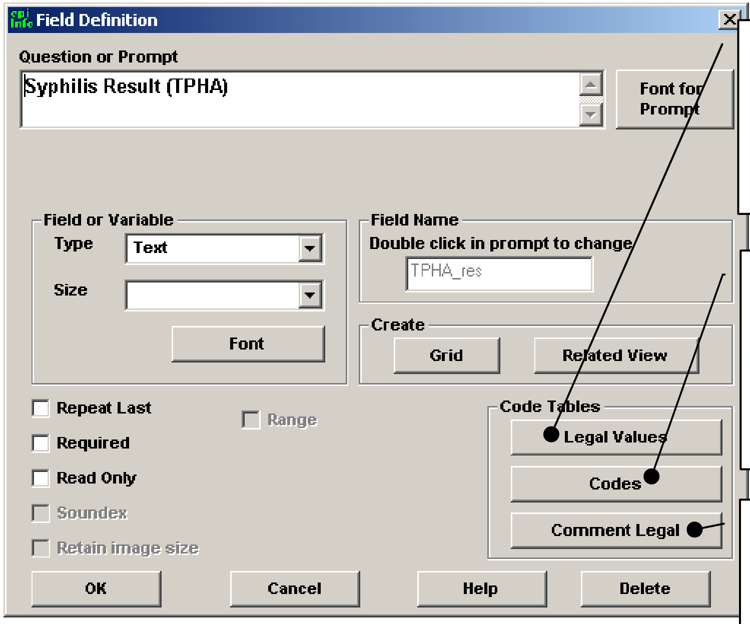
Steps to create code tables

1. Right-click the prompt for *TPHA_res*.

2. Click on the button **Comment Legal,** located in the lower right side of the **Field Definition** dialog box.

3. Click the **CreateNew** button.

*The **Code Tables** provide ways to limit data entry in order to prevent data-entry errors. **Legal Values**, **Codes** and **Comment Legal** values allow you to create a drop-down pick-list of acceptable values with descriptions. The only allowable values are in the Code Tables and are stored in a **Table** in the **Project**. The values may be amended as necessary.*



Field Definition

Question or Prompt: Syphilis Result (TPHA)

Field or Variable — Type: Text

Field Name — Double click in prompt to change: TPHA_res

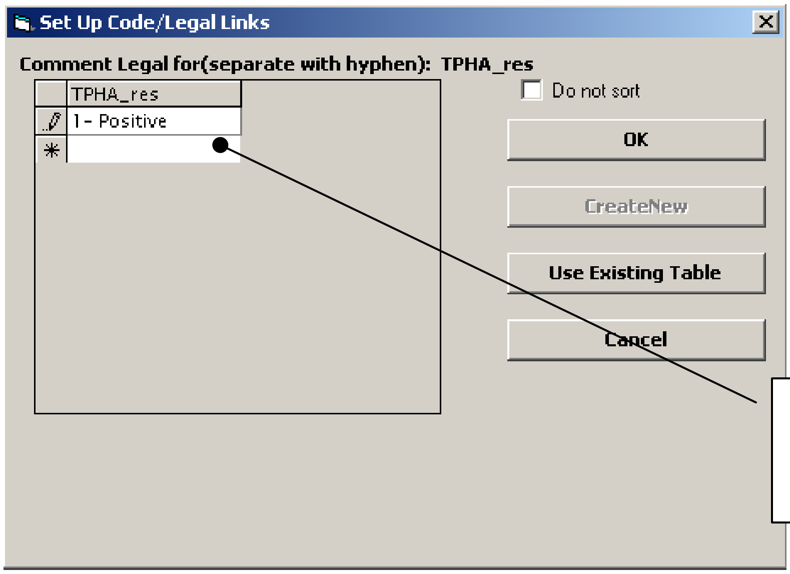
Legal Values tables save the drop-down pick-list value; for example, Marital Status of "Married" appears in the drop-down list, saved as "Married."

Codes tables link values for one variable to values for other variables. For example, selecting Site Name "01" automatically populates values for district ("1") and region ("MVG").

Comment Legal tables save only the values to the left of the dash in the drop-down pick-list. For example, "1" is saved in the data table, although "1-Positive" is displayed on the data entry screen.

Steps to create code tables, continued

The following window should be viewable:



Comment Legal for(separate with hyphen): TPHA_res

TPHA_res
1- Positive

Epi Info displays a table into which you can enter acceptable values for a variable.

4. Enter values for *TPHA_res*:

1–Positive
2–Negative
98–Missing

5. Click the **Do Not Sort** checkbox; otherwise the list will be sorted alphabetically.

***Do Not Sort**, by default, is not checked. When left unchecked, the values entered in a code table will be sorted alphabetically during data entry. When checked, the values will be listed in the order they were entered when creating the Legal Values.*

6. Click **OK** to save the Comment Legal Values.

7. Click **OK** again to save the field properties.

*A Code Table that has been previously created for one variable can be reused by clicking the **Use Existing Table** button instead of typing in new values. For example, if you create a table for the variable mother's occupation, you might reuse that table rather than type the same table again if you were also interested in creating a variable for collecting the father's occupation.*

Moving Fields

Fields should be positioned as similarly as possible to those on the written form.

1. Left-click the **“Syphilis Result (TPHA)”** label.

2. While holding down the left mouse button, move the variable to align it with the other variables.

If you are not able to position the field exactly as shown, you may need to select **Format** on the menu bar, select **Settings**, and uncheck the default setting of **Snap to Grid On**.

Try it yourself!

Activity 3, Move Variables

Move the additional date variables to the appropriate locations, using Appendix D as a guide.

Resizing Fields

To resize a text field that has a codes table linked (a drop-down listing), use the ***Alt+Left click*** *key combination on the data-entry box.*

1. **Alt + Left click** the *Syphilis result (TPHA)* data-entry box (not the label) to make blue dots appear around the data-entry box.

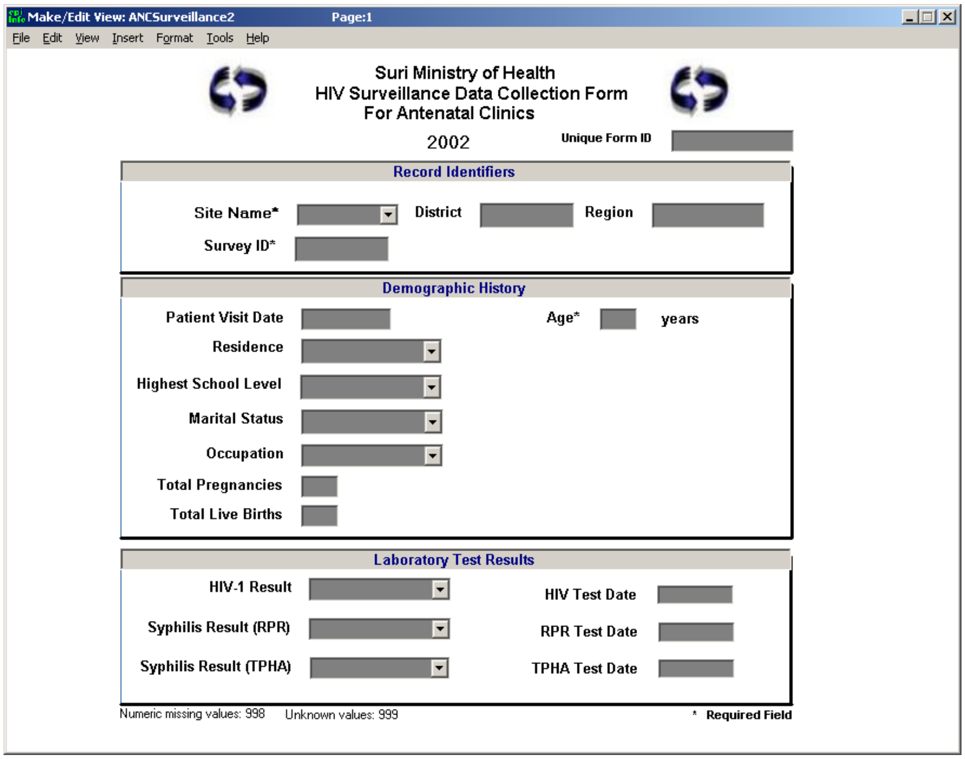
2. Move the mouse cursor over one of the dots. The cursor shape will change to a line with two arrowheads.

3. When the cursor shape has changed, drag it to resize the field.

Only text and multi-line fields are resizable. Number and Date field lengths are determined by the pattern of each field.

Once you have completed moving and resizing fields, your questionnaire should look like the data-entry screen shown below and the data-entry form for 2002 in Appendix D.

Data-entry screen


Make/Edit View: ANCSurveillance2
Page:1
File Edit View Insert Format Tools Help

Suri Ministry of Health
HIV Surveillance Data Collection Form
For Antenatal Clinics
2002

Unique Form ID

Record Identifiers

Site Name* District Region
Survey ID*

Demographic History

Patient Visit Date Age* years
Residence
Highest School Level
Marital Status
Occupation
Total Pregnancies
Total Live Births

Laboratory Test Results

HIV-1 Result HIV Test Date
Syphilis Result (RPR) RPR Test Date
Syphilis Result (TPHA) TPHA Test Date

Numeric missing values: 998 Unknown values: 999 * Required Field

Changing the Tab Order

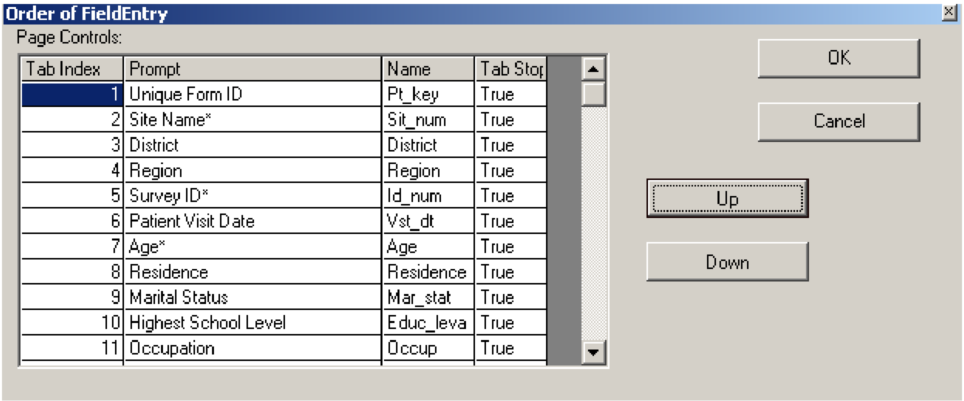
The flow of data entry should be predictable, using a left-to-right, then top-to-bottom tab order, as one would use when filling out a form or reading a page. In Epi Info, variables may not be placed on the view initially in a way that creates a left-to-right, top-to-bottom tab order. To modify the tab order using Epi Info:

Steps to change the tab order

1. Click on the **Edit** menu and then on **Order of Field Entry (Tab Order)** to see the tab order.

2. Click on the prompt you want to be moved and use either the **Up** or **Down** button until the variable is in its proper place in the order.

Steps to change the tab order, continued



Tab Index	Prompt	Name	Tab Stop
1	Unique Form ID	Pt_key	True
2	Site Name*	Sit_num	True
3	District	District	True
4	Region	Region	True
5	Survey ID*	Id_num	True
6	Patient Visit Date	Vst_dt	True
7	Age*	Age	True
8	Residence	Residence	True
9	Marital Status	Mar_stat	True
10	Highest School Level	Educ_leva	True
11	Occupation	Occup	True

3. Repeat the process with other fields until the desired tab sequence is achieved.

4. Click **OK**.

Changing the tab order only changes the order of field entry. It does not physically move the fields on the page to match any changes made with this function.

After the tab order is set, moving a field on the page will modify the tab order to reflect the movement, reordering in a left-to-right, top-to-bottom order.

Try it yourself!

Activity 4, Update the Data Dictionary

Update the Appendix E–Data Dictionary for the Suri 2001 ANC Survey with the additional three syphilis-related variables. Be sure to fill in all columns and make a notation in the Versioning column that these variables were added on today’s date.

Developing Data and Document Storage Strategies

In preparation for beginning a new survey round, it is critical to review your data and document management strategy. Specifically, you should ensure that there is a well-designed file structure layout that clearly specifies the locations of files and their purposes. You may choose to separate databases by year and then further by the type of dataset. Some of the datasets you might generate during the data entry and reporting process include:

- original raw dataset
- dataset of double data-entry reports (to be further cleaned)
- cleaned dataset
- data set for analysis purposes
- current backup of each dataset.

Backup survey data

It is often useful to save these databases separately to avoid overwriting data or deleting important data. It may also be necessary to make a daily backup and keep previous backups for the length of the survey so changes made to the database can be recovered if errors are found. Remember that backup data should be stored in a physically secure place at another building or facility in a commonly available external medium such as a CD-ROM or zip drive.

Try it yourself!

Activity 5, Design an Epi Info Data-Entry Screen

Optional. Using your country's ANC form or a form provided by the trainer, practice designing an Epi Info Data-Entry screen and creating a data dictionary. Be sure to create a new folder to store your project.

Exercise 3
Validating Data Entry

Overview

What this exercise is about

In Exercise 1, some basic rules were set up that governed what type of data would be collected and how it would be collected on the data-entry form. In Exercise 2, new variables were added for the 2002 data-entry screen, including Code Tables to validate data entry. The goal of both exercises is to encourage staff working with the paper form and/or the data-entry system to collect and enter data accurately and more consistently.

The former part of Exercise 3 provides the opportunity to learn more about methods for validating data entry using check code. Simple check code will ensure that routine data-entry errors are eliminated wherever possible. The latter part of Exercise 3 ensures that system documentation appropriately reflects the check code in the system. Internal and external system documentation of check code is another way to ensure that all are aware of how data are being processed during entry.

What you will learn

At the end of the exercise, you will be able to:

- understand the importance of data-entry validation
- use basic commands in Epi Info to validate data entry
- create basic messages to communicate with the user
- document check code in the accompanying system documentation.

Starting location

Epi Info Main Menu

Resources

Appendix F – Check Code and Documentation for the Suri HIV Sentinel Surveillance System

Validating Data Entry Using Check Code in Epi Info

Data-entry staff may inadvertently enter data, such as ages or dates, in error for a variety of reasons. For example, handwriting on forms may be difficult to read or staff may be unfamiliar with the forms and content-specific vocabulary.

In principle, it is always preferable to correct data-entry errors as they occur, rather than having to clean the data after entry. Epi Info provides a simple language-check code that assists in validating data entry. To see the check code that was developed for the 2001 system, and subsequently included in the 2002 system:

Data entry validation steps

1. Click the command button for **Make View** on the Epi Info main page to enter Make View.

2. From the **File** menu, open the file *C:\ANC_Suri\ANC2002\sys02a.mdb*

3. Select the view *ANCSurveillance2*.

   The *C:\ANC_Suri\ANC2002\sys02a.mdb* project file that you are opening in this step is an electronic copy of the 2002 system *sys02.mdb* that you created and saved in ANC2001 in Exercise 2. The *sys02a.mdb* file has been reviewed by the consultants and shown not to contain any known errors or bugs. The file *sys02a.mdb* has been saved in the folder *C:\ANC_Suri\ANC2002* to keep the 2001 and 2002 data files organised separately.

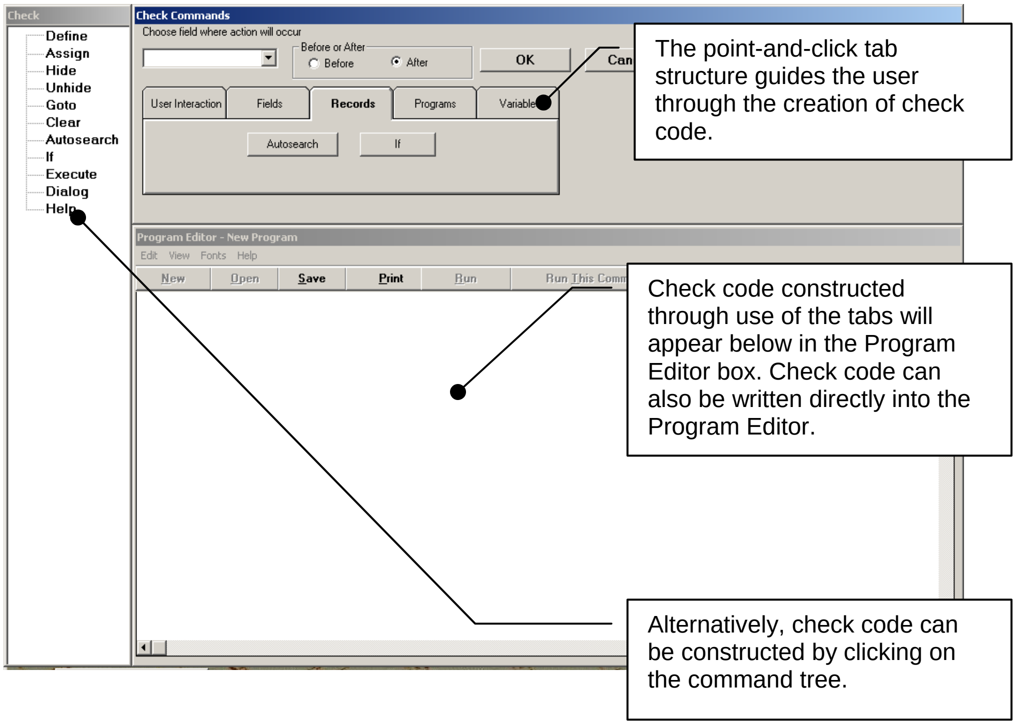
4. Click the **Program** button on the left panel of the window to activate the check code editor.

5. From the drop-down list under **Choose field where action will occur**, select any variable with an asterisk (*) before its name.

The check code for that variable will be displayed in the program editor.

Data-entry validation steps, continued

To familiarise yourself with the check code point-and-click interface using the tab and command tree functions, review the following:



The point-and-click tab structure guides the user through the creation of check code.

Check code constructed through use of the tabs will appear below in the Program Editor box. Check code can also be written directly into the Program Editor.

Alternatively, check code can be constructed by clicking on the command tree.

Tab structure

In the point-and-click environment, it is useful to familiarise yourself with the tab structure, since this is how check code in Epi Info can most easily be created:

- **User Interaction**
  Dialog: Pop-up message windows to make user aware of information
  Help: Opens a new window that directs the user to specific Help information

- **Fields**
  Hide: Hides a field, or fields; assists in preventing anomalies in data entry
  Unhide: Restores a hidden field, or fields, to make them available for data entry
  GoTo: Skips a field, or fields, and directs the cursor to a specific field for data entry
  Clear: Deletes any data entered in a specified field

Tab structure, continued

- **Records**
  Autosearch: A program that searches fields for matches
  If: Provides the ability to create conditional statements

- **Programs**
  Execute: Provides the ability to run additional programs

- **Variables**
  Define: Provides the ability to define new variables
  Assign: Provides the ability to assign values to variables

6. Click **Cancel** to get back to the Make View data-entry screen after reviewing the tab structure.

Using Simple Check Code Commands to Identify Possible Errors

Errors in related variables

The ANC form includes two variables, *par* and *grav*, which have a special relationship that can be tested to identify possible errors during data entry. This relationship can be summarised as follows:

Except for instances of multiple births from a pregnancy (e.g., twins), parity (*par*)—the total number of live births—should never be greater than or equal to gravidity (*grav*)—the total number of pregnancies, including the current pregnancy.

We want to set up check code to test this relationship when entering data for *grav* and *par*.

To set up check code to alert the user of a potential data-entry error when *par* is greater than or equal to *grav*, we will use an If/Then statement. In If/Then statements, if X is true, than Y occurs.

```
If <condition(s)> Then
<statement(s)>

Else                    *optional
<statement(s)>     *optional
End
```

where <condition(s)> is the condition, or conditions, to be met, and <statement(s)> is the check code to execute based on the condition’s truth.

Steps to set up check code

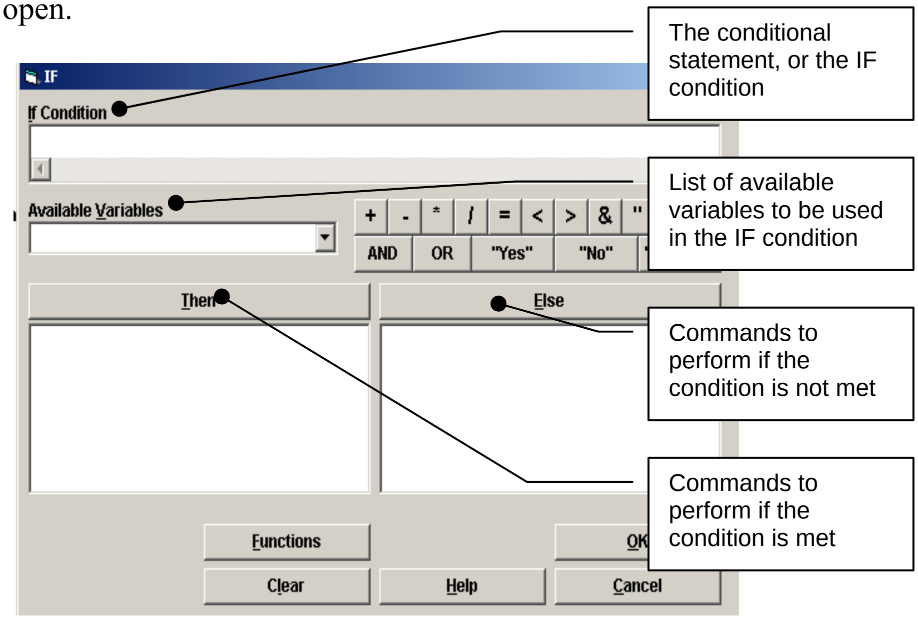
Follow the steps below to set up the check code using If/Then check code:

1. Open the check code editor again by clicking the **Programs** tab.

2. Select the *Par* field from the field list.

3. Select **After**. *Par* can only be validated after data for that field have been entered.

4. Click **If** from the command tree on the left-side panel, or select the **Records** tab and click the **If** command button. The **If** dialog window will open.



The conditional statement, or the IF condition

List of available variables to be used in the IF condition

Commands to perform if the condition is not met

Commands to perform if the condition is met

5. Select *Par* from the **Available Variables** list box as the first argument, then click the > and = buttons. You can also type *Par*>= into the **If condition** box.

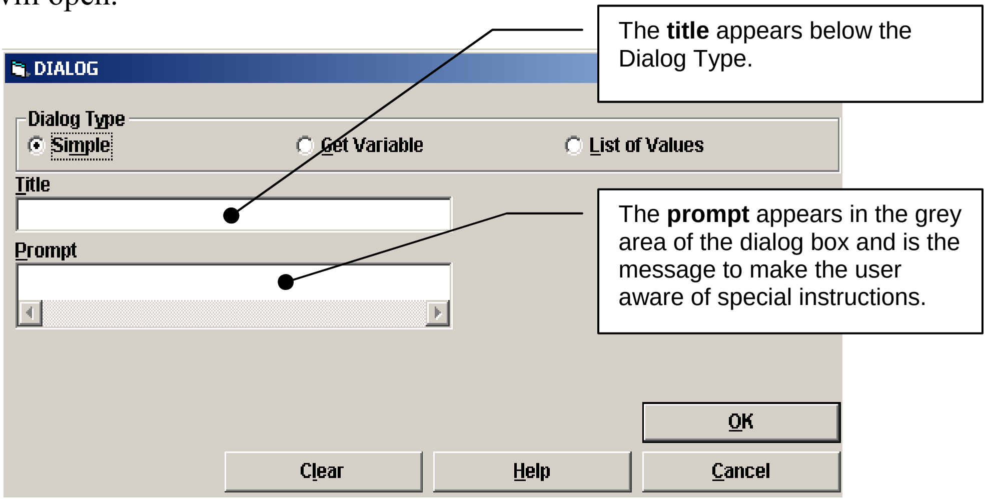
6. Select the variable *Grav* from the drop-down list of **Available Variables**.

7. Click the **Then** command button located below the **Available Variables** list box to construct a statement that will execute if the condition is true. The **Make/Edit View: Check Commands** window will open.

A message or dialog box is often useful to alert the data-entry personnel of a possible error, in addition to performing an action such as clearing the entered value or allowing the entry person to move to the next field.

Steps to set-up check code, continued

8. Select the “User Interaction” tab and click **Dialog**. A **Dialog** window will open.



The **title** appears below the Dialog Type.

The **prompt** appears in the grey area of the dialog box and is the message to make the user aware of special instructions.

9. In the **Title** box, type: *Possible Data Entry Error*

10. In the **Prompt** for the **Dialog** box, type: *Total number of pregnancies is usually greater than the total number of live births.*

11. Click **OK** to exit the **Dialog** box, then **OK** to exit the **If** box.

    The following text should appear in the Program Editor:
    IF Par>=Grav THEN
    DIALOG “Total number of pregnancies is usually greater than the total number of live births” TITLETEXT=“Possible Data Entry Error” END

13. Click **Save** in the Program Editor window.

*Clicking **Save** checks the syntax of the text editor box, then saves it. If there is an error, the code statement(s) will be highlighted and an error message box will appear. To continue, correct the error and click **Save** again.*

14. Click **OK** in the Make/Edit View: **Check Commands** box to exit from the Program Editor window.

At this time, we have not looked at the Enter program. It should become good practice to validate check code as it is created. To do this, click the File menu, and select **Enter Data**. You may be prompted to create a data table. Click “Yes.” (This will be explained in more detail in the next exercise.) Now the check code created can be checked for the expected action.

Using Program Check Codes to Create Skip Patterns

Skip patterns

Often, it is helpful to guide data-entry staff through the process of entering data, allowing them to skip entry of values when appropriate, either by hiding fields or by automatically going to other fields based on an entered value.

Try it yourself!

Activity 1, Hide Data Field

For the field *TPHA_res*, hide the date field if the value of *TPHA_res* is missing (e.g., “98”).

1. Click the **Program** button on the left panel of the window to activate the check code editor.

2. From the drop-down list **Choose field where action will occur**, select the variable *TPHA_res*.

3. Create a check code that will hide *Tpha_Date* if the test result is missing (i.e., value of *TPHA_res* = “98”). Consider the alternative, and add an ELSE condition to unhide *Tpha_Date* if the test result value should be changed.

4. Modify check code for the fields *HIV_res* and *RPR_res* with an ELSE condition.

Activity 2, Create Check Code to Control Entry Date

For the fields *Tpha_date and Rpr_date*, create check code that ensures that the entry dates of the tests are between 01/01/2002 and 31/12/2002, inclusively.

Hint: If you need assistance in writing the check code, refer to the field *HIV_date*.

When dates are evaluated in check code, the format must be in MM/DD/YYYY, although you may set the date format to be entered in DD/MM/YYYY.

Developing Complex Check Code

Rules for validating age

The ANC sentinel surveillance form and data-entry screen are relatively simple when it comes to check code, relying primarily on code tables, legal values and range checks. Age is an important variable for analysis. Consequently, to ensure that *Age* is entered accurately, several rules have been created to validate the values entered for this field. These rules are as follows:

- No woman should be included in the system unless she is aged 12–49.
- *Age* is a required field, but may be entered as missing or unknown using the appropriate codes, 998 or 999.

Try it yourself!

Activity 3, Develop Check Code for Age

1. Based on the rules stated for *Age*, in the space below, write check code that will enforce those rules and will alert the user of a possible problem when an inappropriate value is entered. Consider the limitations that we want for age and the properties for the field.

2. Using the Make View Check Code Program, create the logical check code for *Age*.

3. Compare your answers for *Age* to those in Appendix I. You will be able to validate the check code in Exercise 4.

Documenting System Check Code in the Program Editor Window

Check code comments

System check code is critical to document in a single location so that at a glance, users of your system and the data are aware of the rules guiding data entry. For this reason, it is critical to provide a single document that captures check code by variable in detail. To assist users in understanding your check code, it is often helpful to provide comments within the check code. Check code comments can be included in the program editor by beginning the comment line with an asterisk (*).

Hint: Two options exist for adding comments:

- Enter the asterisk and type the necessary comments. Even if your comment is a multiple-line comment, do not press the ENTER key at the end of the line. Only press ENTER when you have completed your comments. This tells Epi Info to look for a command next.
- Enter an asterisk before each line that you wish to be a comment.

Try it yourself!

Activity 4, Document Program Code

In the Check Code program, document the check code for *Age* and *TPHA_res* so that non-programmers will understand the assumptions made and the steps in the code that allow each assumption to be met. When done, close Check Code and Make View.

Documenting System Check Code in an Outside Source

System check code should also be documented in a word processing file to facilitate review by non-programmers. Except for the most recent additions, documentation of check code for the 2002 ANC system can be seen in Appendix F – Check code and documentation for the Suri HIV Surveillance System. Columns are included that reference the following:

- the specific variable where the check code occurs
- the check code itself with documentation for non-programmers
- the trigger action indicating when the check code is performed
- additional variables that are referenced in the check code, if any.

Try it yourself!

Activity 5, Complete Check Code and Documentation

In Appendix F – Check code and documentation for the Suri HIV Surveillance System and fill out the information for *par*, *TPHA_res, Tpha_date, Rpr_date,* and *Age*.

Exercise 4

Overseeing and Performing Data Entry

Overview

What this exercise is about

With a complete but untested system ready for data entry created by the Surveillance Team and the consultants, the 2002 sentinel surveillance round commences. Until data-entry staff can be hired, you have been asked to test the properties of the views and the check code that were completed in Exercise 3 by entering the incoming 2002 forms.

Once you have entered the first batch of reports from Banket, you will also be asked to search and find specific records using the Epi Info Find feature.

What you will learn

At the end of the exercise, you will be able to complete the following tasks:

- enter additional records
- navigate through Epi Info records in the Enter field
- search and find specific records.

Starting location

Enter Data, C:\ANC_Suri\ANC2002\sys02b.mdb

Resources

Appendix G – Round 3 – Year 2002 Data-Entry forms (6 Banket forms)

Entering Data into Epi Info

With edit checks completed, accurate data entry should be simple in Epi Info.

In Epi Info, the **Enter Data** program is used to enter and save data. It is a separate tool from **Make View**, but uses the outputs (namely, the data-entry screen and check code) to allow data-entry staff to enter data and search for records. The advantage of using the **Enter Data** program in Epi Info is that no one, including data-entry staff, can modify or change the data-entry screen and its properties.

**Steps to
enter data**

To access Epi Info **Enter Data** and begin to enter forms from the 2002 survey:

1. From the Epi Info main menu, click on the **Programs** menu.

2. Click on **Enter Data**. The program opens.

3. From the **Enter Data** program, click on **File**.

4. Click on **Open...**

5. Navigate to C:\ANC_Suri\ANC2002\sys02b.mdb.

6. Click **Open....**

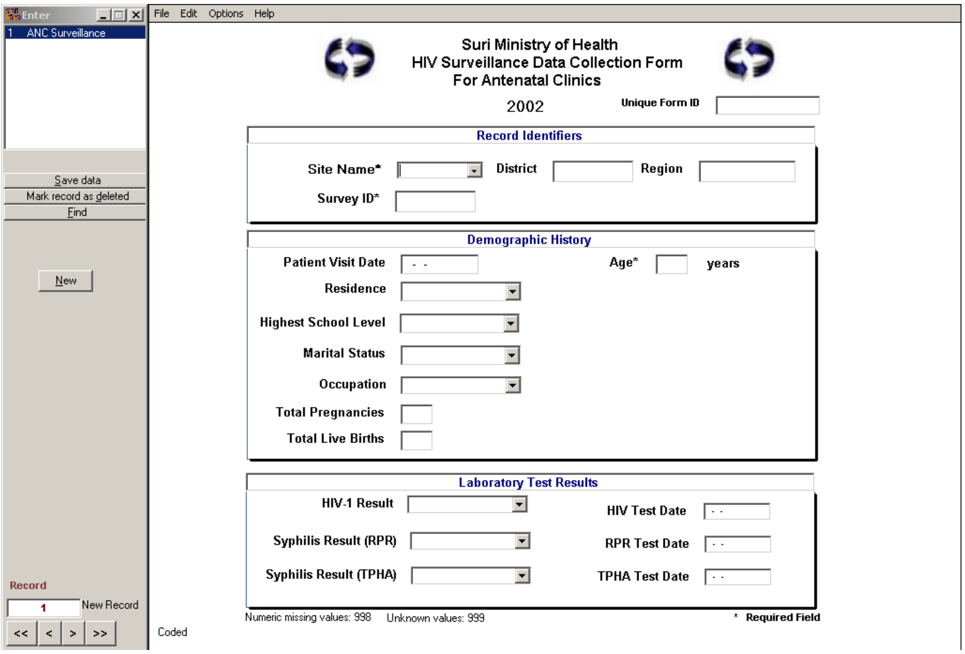
7. Select the view *ANCSurveillance2*.

8. Click **OK**.

The *C:\ANC_Suri\ANC2002\sys02b.mdb* project file that you are opening in this step is an electronic copy of the 2002 system *sys02a.mdb* system to which you added check code in Exercise 3. The *sys02b.mdb* file has been reviewed by the consultants and shown to not contain any errors or bugs.

Steps to enter data, continued

The 2002 HIV Surveillance Data-Entry screen for Antenatal Clinics should appear as follows:



Suri Ministry of Health
HIV Surveillance Data Collection Form
For Antenatal Clinics

2002 Unique Form ID

Record Identifiers

Site Name* District Region

Survey ID*

Demographic History

Patient Visit Date - - Age* years

Residence

Highest School Level

Marital Status

Occupation

Total Pregnancies

Total Live Births

Laboratory Test Results

HIV-1 Result HIV Test Date - -

Syphilis Result (RPR) RPR Test Date - -

Syphilis Result (TPHA) TPHA Test Date - -

Numeric missing values: 998 Unknown values: 999 * Required Field

When entering dates, you may enter the 2-digit day, 2-digit month and 2-digit year. It is not necessary to type the 4-digit year. The current year set in the Window's system date will be assumed unless you type in a different year value.

For fields created as legal or comment legal values, in some instances, typing the first character will automatically populate the field with the proper response. There may be times when two or more characters are necessary; for example, when you have two values that start with "A" and you want to select the second value.

9. Enter the first form exactly as it appears in Appendix G into the *ANCSurveillance2* view.

Epi Info requires that 'must enter' fields, such as Age, are completed before moving to the next record or before exiting the application. For this reason, use the required field's checkbox only when you give the data-entry staff specific directions for how to respond if there is no response or the response is not legible.

Try it yourself!

Activity 1, Enter and Save Data

1. After entering the first form, click the **New** button (located in the tree command structure on the left side of the data-entry screen) to create the next empty record if you did not already press enter on the last field of Record 1.

2. Enter the five additional forms exactly as they appear. Missing responses should be considered "Missing." If you identify any potential errors in the collection of the data or are unsure of how to enter a response in the system, make a note of these anomalies on the side of the form by the variable.

3. Click the **Save Data** button in the tree command structure on the left side of the data-entry screen.

It is not necessary to save data before exiting or navigating through records. However, it is good practice.

Navigating Through and Finding Records

Steps to find records

1. On the lower left-hand side, under the record counter, click the arrows to navigate the entered records.

The << sign brings the data-entry screen to the first entered record, while the >> sign brings the data-entry screen to the last entered record.

The < brings the data-entry screen to the previous record, and the > brings the data-entry screen to the next record.

To navigate to a specific record, click in the white box and highlight the current record number, type in the desired record number, and press the Enter key.

2. To find a record, click the **Find** button on the left-hand side. A **Find Record** screen appears with a list of all available fields.

3. To test the capabilities of the Find, click the *Age* field to be prompted with a blank field.

4. Type *29*.

Steps to find records, continued

5. Click **OK** or press **Enter** on the keyboard.

6. Depending on how you interpreted the ages on the form, one to two records should appear. Double-click the row indicator (the grey area to the left of one of the records), to bring it to the data-entry screen.

7. Once you have pulled up the form and reviewed, click **Find** again to go back to the results of your search. Click **Reset.**

*The Find also has the ability to search **wildcard**. For example, typing 00* (asterisk) in the Id_num field will return all files with 00 in their Id_num field.*

8. To test the capabilities of the Find to identify a specific record by ID Number, click the *Id_num* field to be prompted with a blank field.

9. Type "003."

10. Click **OK** or press **Enter**.

*Up to six fields can be selected to perform a Find. Selecting multiple variables works like a conditional AND, returning only those records that meet the conditions of all of the variables. To select the search fields, click the desired fields. To deselect a field, click the selected field again in the "**Choose Search Field**" box. Clicking or adding fields after you've begun selecting multiple conditions and entering search criteria will erase the contents of the fields already containing criteria.*

Try it yourself!

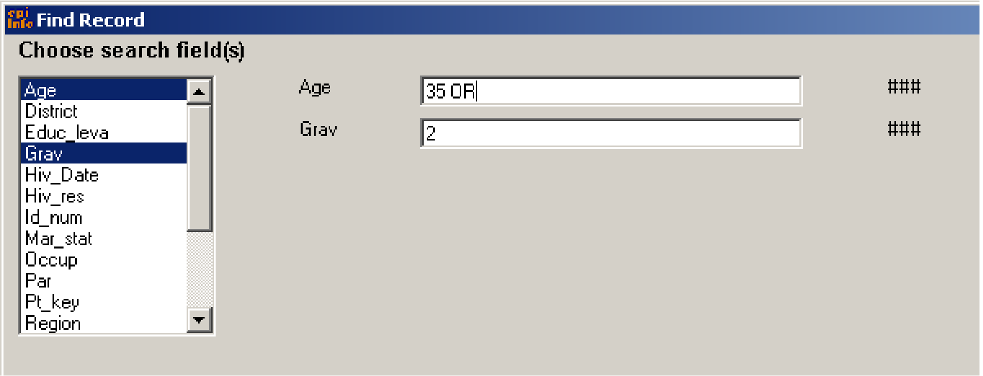
Activity 2, Identify Survey ID Number

Identify the Survey ID Number where the Site Number is 01, the patient visit date occurred on 24/06/2002 and the woman had no previous births.

Write your answer here:

The Find functionality automatically executes an AND condition when multiple fields are included in a search. To create an OR condition, the word OR has to be explicitly placed after the field condition. See example below:

Age = 35 OR Grav = 2



Find Record

Choose search field(s)

Age: 35 OR

Grav: 2

Exercise 5

Developing and Documenting Data Cleaning

Overview

What this exercise is about

In Exercise 4, you entered six records, noting some obvious and some questionable data collection errors written on the forms. At the same time, it is possible you introduced additional errors as you entered the data. To prevent data-entry errors from remaining in the file for analysis, the team should have a well-defined data-cleaning plan that systematically:

- outlines a process for identifying possible errors, how and by whom they should be resolved and in what time period
- identifies specific anomalous values (i.e., values out of range or unexpected) or errors in the database
- documents for historical reference changes to the database to correct the error on the basis of this review process.

In Exercise 5, you will develop this plan and begin to operationalise it. One of the first steps will be to double-data enter the six reports received by the MoH, compare the files, and document possible errors and their resolution in a data-entry audit log. The remainder of the data cleaning plan and documentation of changes will be completed in Exercise 6, once the approximately 6 000 report forms are received at the Ministry of Health.

What you will learn

At the end of the exercise, you will be able to:

- design and carry out a plan for cleaning data, including identifying and resolving errors
- perform double data entry and compare records to resolve differences
- fill in a sample data-entry audit log.

Starting location

Enter Data, C:\ANC_Suri\ANC2002\sys02bdde.mdb

Resources

Appendix G – Round 3 – Year 2002 Data-Entry forms (6 Banket forms)
Appendix H – HIV Surveillance Data-Entry Audit Log

Developing a Data Cleaning Plan

Regardless of how carefully data collectors fill out forms or how comprehensively check codes are used in the system, errors or anomalies in the data may still occur. As the team providing oversight to the survey, it is your job to identify these errors and anomalies as soon as possible and to attempt to resolve them systematically and consistently throughout the survey. Having a written data cleaning plan to which all stakeholders agree will ensure that errors are consistently addressed in a timely fashion. When you are developing a data cleaning plan, it is useful to address the issues starting on the next page.

Step 1: Develop a process

1. Outline a systematic process for identifying possible errors, how and by whom they should be resolved, and in what time period.

As part of your data cleaning plan, you should begin by identifying possible errors immediately evident on the form. For example, multiple responses may be marked in a question or a response may be unreadable. In both of these situations, it is clear that there is a possibility of introducing an error into the database.

Next, you should identify those errors that can be detected during data entry according to pre-established check code. For example, a response on the form not falling into the set of pre-defined responses developed using Comment Legal values may be considered an error.

Finally, you may need to explicitly develop a process that identifies errors that don't result from a specific form, but are introduced during data entry as a pattern of erroneous responses at a particular site, by a particular data-entry staff or in a particular variable across all sites.

- As an example, some sites may not ask clients about the total number of previous live births and instead will mark zero every time. While this value is allowable, it may not be correct.

Step 1: Develop a process, continued

- Another example is finding that all HIV test results on a particular day were positive. While it could be that all tests were truly positive on this day, it could also mean that the samples during that day were contaminated by one positive sample or that the technician was unfamiliar with how to perform the lab test.
- Finally, data-entry staff may simply type one value when they should have typed another.

These errors, although not evident on the form or identified during data entry, are also critical to identify during data cleaning.

Once the possible errors have been identified, a process is needed to systematically and consistently resolve them. As part of this process, the person(s) responsible for resolving the possible error and the time period in which the resolution will occur should be specified. In the case above where values are illegible, the data manager in the MoH could follow up with the site supervisor to determine if other information exists to clarify the response. If no information is obtained within a week, the value will be considered “missing” in the database as determined by the data manager.

Once the process for identifying and resolving anomalies or errors has been documented, the data-entry clerk and other staff overseeing data management should receive a list of these rules to reference. In all cases, be sure to instruct staff on how to flag possible errors consistently, by noting these either on the form or in a 'problem' log.

Step 2: Write a list of steps

2. Write a list of steps for identifying data anomalies or possible errors that are clear and specific to the application and then translate these steps into computer code where needed.

Steps for cleaning data should start with identifying the most obvious and immediate errors.

- Review completed forms centrally as they are received and prior to data entry to resolve errors (e.g., review for missing data including clinic or site name).
- Use check code during data entry to highlight potential errors in the completed form (e.g., out-of-range ages).

Step 2: Write a list of steps, continued

- Conduct double data entry of the forms (i.e., comparison of the same form entered by two different staff members) to identify potential data-entry errors or differences in interpretation of the completed form. When you have large files and limited resources, you may want to enter only a sample of records. However, in most situations, it is worth the required resources to enter the forms again and compare them to detect errors.
- Generate simple lists and frequencies to identify anomalies of responses (e.g., a high number of missing values at a site, incorrectly entered text variables, dates that appear to be outside the survey range and testing dates that occur before a client's visit).

Step 3: Document errors

3. Document the errors and their resolution in a data-entry audit log.

Once a process for identifying errors has been established and there is agreement on how to resolve them, it is important to use a data-entry audit log to record errors, the method of resolution and the resolution itself.

A data-entry audit log is an electronic or written record of changes to the data that were made as a result of the data-cleaning process. The audit log is an important document, in that it ensures that data-cleaning decisions are consistently carried out over time. In addition, it can serve as a historical document or archive detailing what decisions were made and what actions were taken to the database.

When errors are found and changes are made to the data, a record of each transaction should be made in the audit log. Keeping notes helps you to ensure that everyone is working from the most recent cleaned file. Every entry in the log should be dated and initialed by the person who made the change.

Step 3: Document errors, continued

What an audit log includes

A sample data-entry audit log might contain the following fields:

- Date – the date that the possible error was identified
- Survey Site Name (if using form)
- Survey ID Number (if using form) – this, in combination with the survey site name, can uniquely identify a form
- Unique ID Number (if electronic entry) – this can be used to uniquely identify a record
- Variable name and value – the name of the variable (or variables if they are linked) that needs to be clarified, and the current value
- Description of anomaly – a description of the possible error and how it will be resolved
- Resolution made – a description of the final data point entered into the database and how the resolution was made (e.g., the site co-ordinator was called and original log books showed that forms were mislabeled)
- Date of final resolution
- Initials of supervisor or person overseeing the change.

Data-entry audit logs can be created and managed in any word processing or spreadsheet package. Appendix H provides a sample data-entry audit form for your use in completing Exercises 5, 6 and 7.

Try it yourself!

Activity 1, Create a Data Cleaning Plan

Create a written data-cleaning plan for the 2002 ANC survey that includes the following:

1. Rules for:

   - when and how missing variables should be entered
   - when and how missing dates should be entered
   - in what instance and how unreadable values should be entered
   - in what instance and how variables with multiple responses or responses not in the list of allowable responses should be included.

2. A list of steps for identifying data anomalies or possible errors.

Address the methods for doing double data entry, including how many reports should be entered, who should enter them and when they should be entered. In addition, review the data-entry screen and identify which variables should be analysed for:

- missing or unknown values
- outliers
- inconsistencies.

For this third round (2002), the Surveillance Team should receive approximately 6 000 reports. Two data-entry staff should be available to fully support data-cleaning activities.

Performing Double Data Entry

In this section of the exercise, you will work on the first stage of the data cleaning plan by doing double data entry for the first six forms. In Exercises 6 and 7, you will use Epi Info Analysis to do simple data cleaning, to correct errors and to modify records using Enter Data and Analysis.

With only six forms, double data entry of all forms is simple. To perform double data entry for this exercise, move to another team member's computer (or if doing these exercises alone, stay at your own computer), and follow the steps below:

Steps for double data entry

1. Click on the Epi Info Program menu drop-down box.

2. Click on **Enter Data**.

3. Click on **File.**

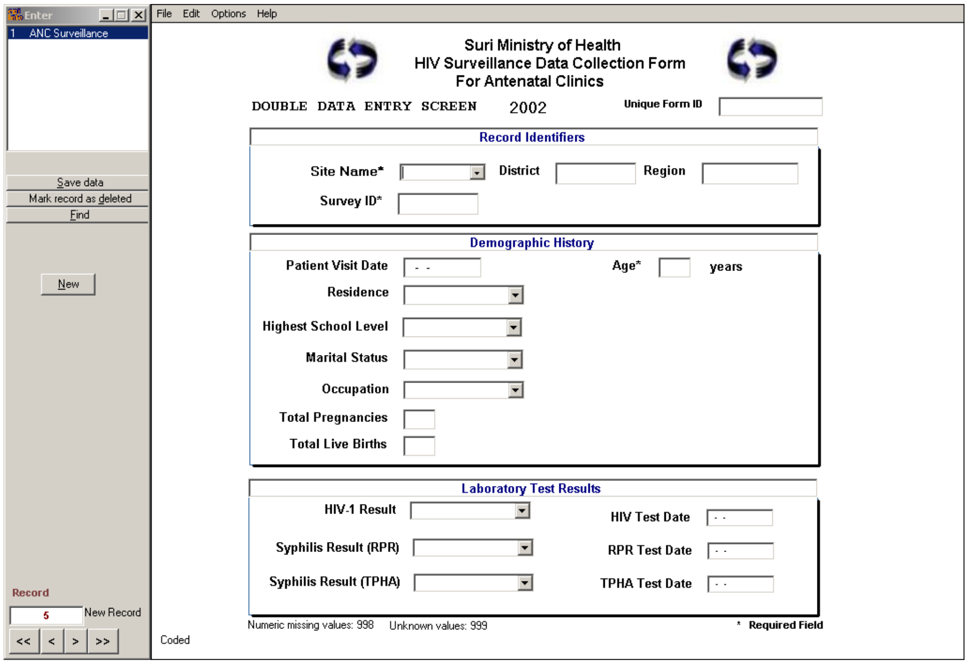
4. Click on **Open...**

5. Select or type *C:\ANC_Suri\ANC2002\sys02bdde.mdb* and click **OK**.

This file is the same structure as the 2002 ANC system that was created in previous exercises; however, an additional text variable has been added to let the data entry operator know that form entry will be done in the Double Data Entry database rather than the primary database.

6. Select the view *ANCSurveillance2*. The 2002 HIV Surveillance Double Data Entry Screen for Antenatal Clinics will appear:

Steps for double data entry, continued



7. Enter records 1-6 from Appendix G.

8. Note in the margins on the form any potential data collection errors or areas where supervisor review would be appropriate.

9. Exit **Enter**. If you have moved to another participant's desk, return to your original location.

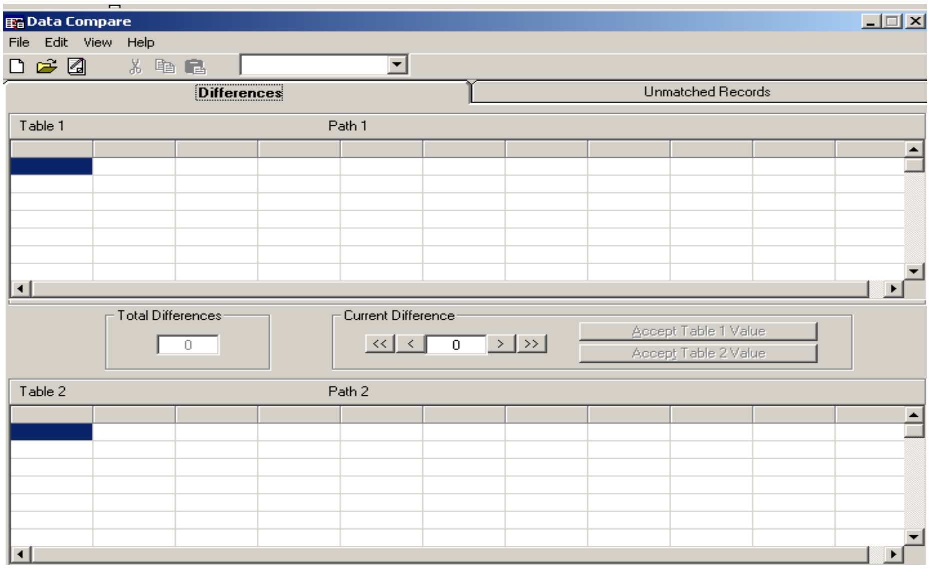
Comparing Data Entered Into the First and Second Databases

Epi Info uses the Data Compare application for finding the differences between two tables or datasets. In this section of the exercise, we will identify differences between the two datasets that were entered: *sys02b.mdb* and *sys02bdde.mdb*.

Steps to compare data

1. From the Epi Info main menu, click **Utilities** from the main menu bar.

2. Click on **Data Compare** to see the following screen:



3. Click **File** from the menu bar.

4. Select **New Script** from the drop-down list. The **Data Compare Wizard** screen will appear.

5. Select **Standard Table** in the Type of Tables option prompt.

6. Click the button with the three dots to the right of the MDB 1 prompt. Navigate to *C:\ANC_Suri\ANC2002\sys02b.mdb*.

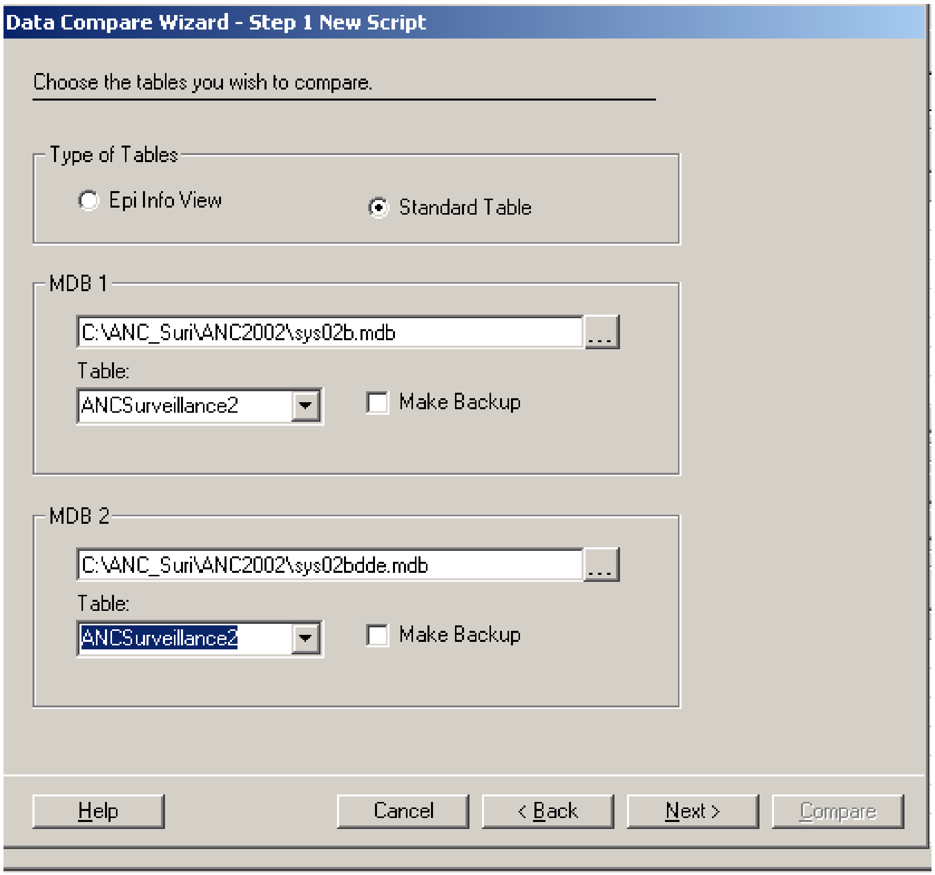
Steps to compare data, continued

7. Choose *ANCSurveillance2* from the drop-down list in the **Table:** prompt.

   The *ANCSurveillance2* contains data from the six forms.

8. Below MDB 2, click the button with the three dots to the right of the prompt. Navigate to *C:\ANC_Suri\ANC2002\sys02bdde.mdb*.

9. Choose *ANCSurveillance2* from the drop-down list in the **Table:** prompt. This table contains six forms that were entered during double data entry.



Data Compare Wizard - Step 1 New Script

Choose the tables you wish to compare.

Type of Tables

○ Epi Info View ● Standard Table

MDB 1

C:\ANC_Suri\ANC2002\sys02b.mdb

Table: ANCSurveillance2 □ Make Backup

MDB 2

C:\ANC_Suri\ANC2002\sys02bdde.mdb

Table: ANCSurveillance2 □ Make Backup

10. Click **Next** to proceed to Step 2 of the wizard.

    Data Compare checks the table structures to ensure the variable names and types are the same.

Steps to compare data, continued

11. Click **Next** to proceed to Step 3 of the wizard.

12. Click the checkbox for the *pt_key* field, which is the unique variable that represents each record in the database.

The variable will be used to match records from the two different databases.

*Data Compare requires a **unique** variable to compare records in two separate databases. If no such uniqueness exists despite the design of a unique key, it is possible that there are multiple records in one or both of the databases with the same identifier key. An error message stating that the selected variable is not unique may mean that you need to review entries in Enter to determine if mis-keying of the unique identifier key has occurred.*

13. Click **Next** to proceed to Step 4 to choose the variables that will be compared in each of the datasets.

14. All of the variables are checked by default. Click **Next**.

15. Click **Next** to skip Step 5. We will not be creating an HTML file. The results are more easily viewed on the screen.

The process of moving through the Data Compare Wizard creates a program called a script that can be saved for future use. The script can be opened and run automatically rather than manually walking through each of the steps as listed above. If you wish to save the script, you can do so at this point.

16. Click the **Save As** button and navigate to the "C:\ANC_Suri\Programs" folder. Save the script with the name "DataComp02.txt". Click **Save**.

17. Click **Compare**.

Data comparison results in Epi Info

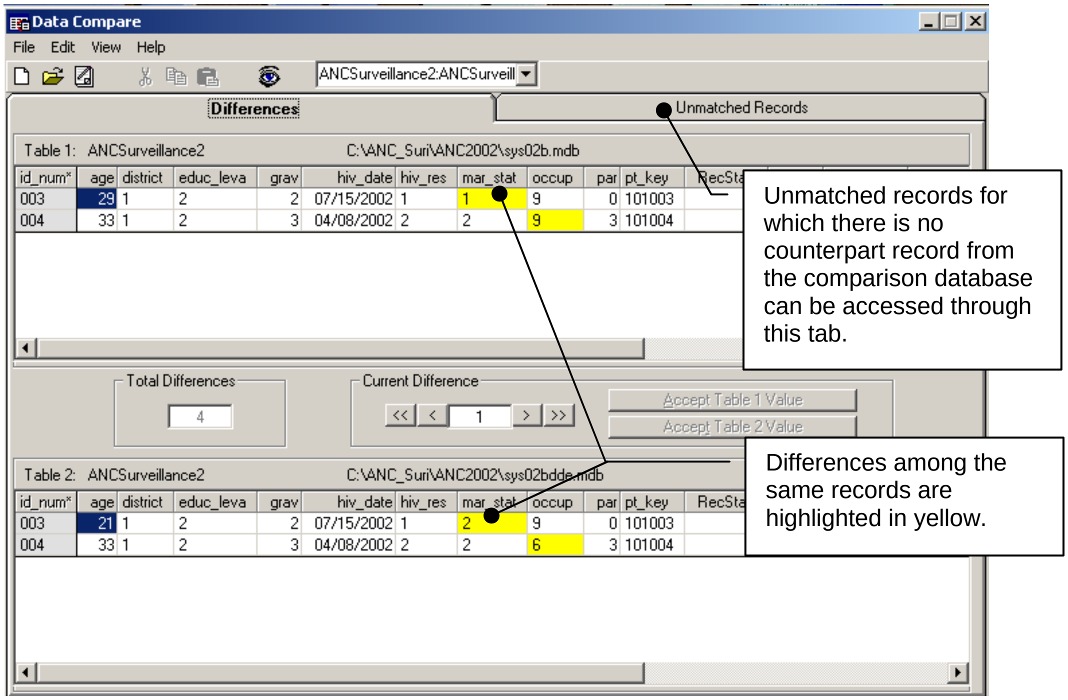


Table 1: ANCSurveillance2 C:\ANC_Suri\ANC2002\sys02b.mdb

id_num*	age	district	educ_leva	grav	hiv_date	hiv_res	mar_stat	occup	par	pt_key	RecSta
003	29	1	2	2	07/15/2002	1	1	9	0	101003	
004	33	1	2	3	04/08/2002	2	2	9	3	101004	

Table 2: ANCSurveillance2 C:\ANC_Suri\ANC2002\sys02bdde.mdb

id_num*	age	district	educ_leva	grav	hiv_date	hiv_res	mar_stat	occup	par	pt_key	RecSta
003	21	1	2	2	07/15/2002	1	2	9	0	101003	
004	33	1	2	3	04/08/2002	2	2	6	3	101004	

Unmatched records for which there is no counterpart record from the comparison database can be accessed through this tab.

Differences among the same records are highlighted in yellow.

If you have no differences, a watermark picture will appear, indicating that there are no differences among the records. If you have any differences that resulted from clear mis-keying, Epi Info will highlight those differences in yellow.

Try it yourself!

Activity 2, Document Possible Errors

Use Appendix H – HIV Surveillance Data Audit Log to document possible errors identified in Data Compare. Be sure to complete Columns A through G, including determining the resolution that should be made in the database according to the process you outlined in Activity 1. For the purposes of this exercise, you can assume that no additional information is available from the clinic site.

Resolving Differences Using Data Compare

By default, Epi Info does not allow you to update data in Data Compare. To change the default setting to make the changes you identified above:

1. Select the menu options under **View.**

2. Click on the **View as Read Only** option to uncheck.

3. This will activate the greyed-out buttons between the two files indicating whether the Table 1 value or Table 2 value should be accepted.

4. Click on either the **Accept Table 1 Value** or **Accept Table 2 Value** command buttons according to your resolution in the data-entry audit log.

Try it yourself!

Activity 3, Use Data Compare to Resolve Differences

Use Data Compare to resolve the remaining differences in the two data files. Be sure to update Appendix H – HIV Surveillance Data Audit Log columns H and I.

Just because you *can* do something in Epi Info doesn't mean you always *should!* In Exercise 7 you will learn more about the drawbacks of editing the raw data file directly from Data Compare and better approaches to resolving differences once they have been identified.

Notes

Exercise 6
Conducting Simple Exploratory Analysis For Data-Cleaning Purposes

Overview

What this exercise is about

Since your initial entry of the six forms, an additional 6 925 forms have been added to the database by data-entry clerks, for a total of 6 931 records in the 2002 round. While oversight by the team during this process was adequate, possible errors may have gone unrecognised in the database. Following the data-cleaning plan below, we will read (open) the database *C:\ANC_Suri\ANC2002\sys02c.mdb*, sort, select, list and perform frequencies of the 6 931 records in the 2002 data set to identify any possible errors or anomalies.

What you will learn

At the end of the exercise, you will be able to:

- read and write Epi Info databases
- use Epi Info Analysis functions including select, list, frequency and table commands to identify data errors and anomalies.

Starting location

Analysis, C:\ANC_Suri\ANC2002\sys02c.mdb: ANCSurveillance2

Resources

Appendix H – HIV Surveillance Data Audit Log
Appendix I – Selected original data-entry forms for 2002 sites

Conducting Simple Exploratory Analysis to Detect Possible Errors

In Exercise 5, the team developed a data-cleaning plan to identify and resolve errors. Part of this plan undoubtedly included simple exploratory analyses, such as generating frequencies and looking for consistencies in dates or validating variables that have relationships. Simple exploratory analysis is a key tool in the data-cleaning plan for detecting remaining errors or anomalies in your database.

For the purposes of this exercise, the simple exploratory analysis section of the data-cleaning plan for the 2002 data set is as follows:

Steps for simple exploratory analysis

1. Conduct simple frequency analyses to check for outliers, anomalies or inconsistencies in the data.

   a. **Frequency of Age** – Guidelines for age stipulate that no age should be less than 12 or greater than 49 and there should be no missing age values, unless indicated by 998 or 999. However, all records with age=12 should have the age variable validated against the form again, because this population of young adults is of particular interest.

   b. **Frequency of Site** – Guidelines stipulate that each site meet the minimum sample size of 300. However, the three large urban sites, “12,” “16,” and “19” each had a sample of 500 or more. These sites were selected to get better statistical precision in calculating prevalence amongst youth aged 12-29.

   c. **Frequency of Gravidity** – Check that all women have at least one pregnancy listed. If not, they should be excluded from the survey since they do not meet eligibility criteria.

   d. **Frequency of Parity**

   e. **Frequency of Syphilis results**

   f. **Frequency of HIV results**

Steps for simple exploratory analysis, continued

2. Perform table analyses to check for consistency of parity and gravidity. Gravidity should always be greater than parity, except in the case of twins.

3. Perform table analyses of the key outcome variable, the HIV test result (*HIV_res*), by site to see if any sites have an unusually high or low HIV prevalence. This may indicate a problem in sample analysis, data collection or data entry that needs to be resolved.

4. Check consistency of dates. Client-visit dates should never occur after HIV and syphilis test dates.

You may have identified other exploratory analyses in the data-cleaning plan that, time permitting, can be further investigated at the end of the exercise. For example, it is often useful to look at frequencies of all variables by site to identify possible problematic data collection patterns. In addition, it is often worth looking at laboratory results by testing day to crudely assess quality.

Using Epi Info Analysis to Read Epi Info Data

To begin conducting simple exploratory analysis, we will first read, or open, the *sys02c.mdb file::ANCSurveillance2* data table using Epi Info Analysis.

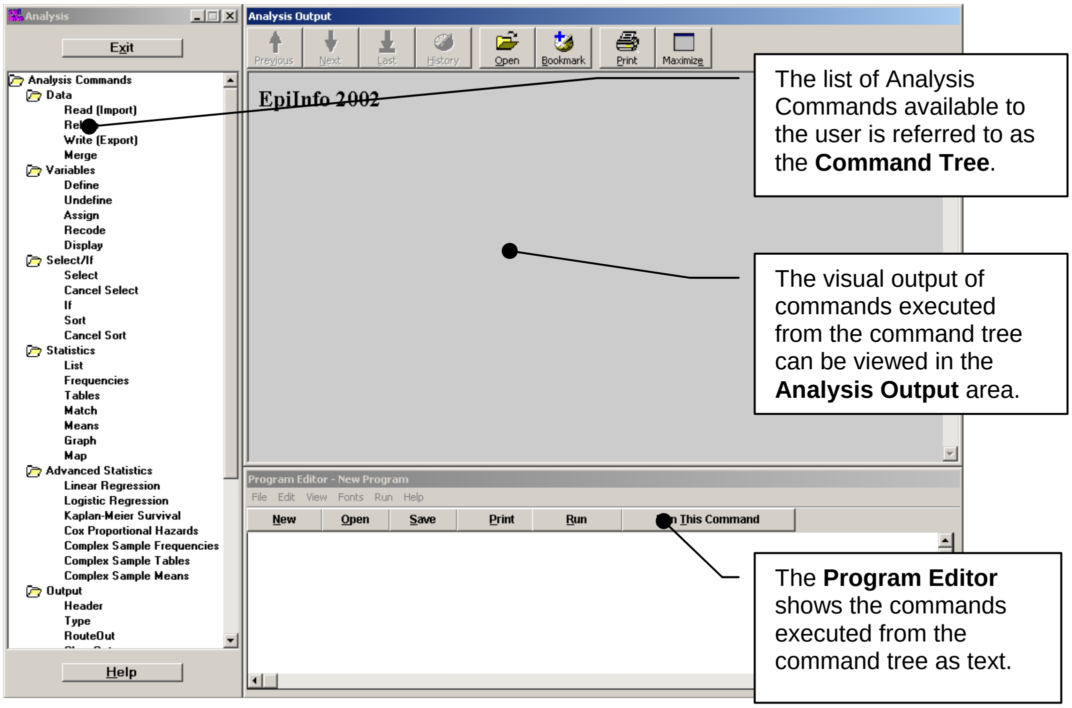
Epi Info's Analysis program can be used to:

- read (i.e., open) data from Epi Info and other database types (e.g., Excel, Access, Epi 6, dbf, etc)
- manipulate and clean individual records or recordsets
- conduct simple and complex statistical data analysis, graphing and mapping.

Reading the 2002 data

To read the 2002 data:

1. From the main Epi Info menu, click **Analyze Data** to access **Analysis**. The **Analysis** program will appear.



The list of Analysis Commands available to the user is referred to as the **Command Tree**.

The visual output of commands executed from the command tree can be viewed in the **Analysis Output** area.

The **Program Editor** shows the commands executed from the command tree as text.

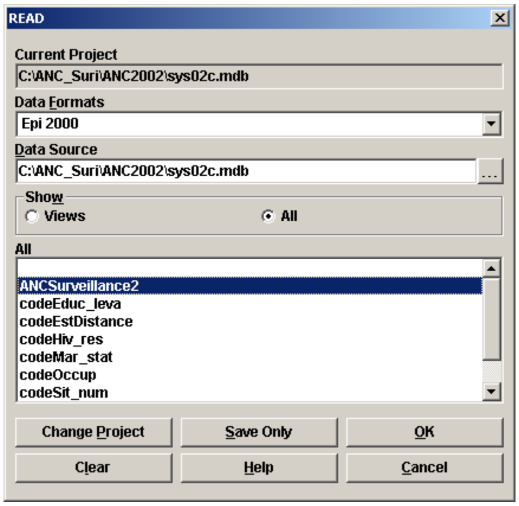
2. Click on **Read (Import)** under the Data folder in the command tree. A dialog window opens.

3. Click the **Change Project** button at the bottom left of the dialog window.

4. Find and select "*C:\ANC_Suri\ANC2002\sys02c.mdb*". Click **Open...**.

Reading the 2002 data, continued

5. Select the **All** radio button to see *ANCSurveillance2*.



READ

Current Project: C:\ANC_Suri\ANC2002\sys02c.mdb

Data Formats: Epi 2000

Data Source: C:\ANC_Suri\ANC2002\sys02c.mdb

Show: ○ Views ● All

All:

ANCSurveillance2
codeEduc_leva
codeEstDistance
codeHiv_res
codeMar_stat
codeOccup
codeSit_num

6. Click **OK**.

Analysis output

The **Analysis Output** area should show the following text:

Current View: *C:\ANC_Suri\ANC2002\sys02c.mdb: ANCSurveillance2*
Record Count: *6931 (Deleted records excluded)*
Date: *8/01/2003 11:09:18 AM*

You have now completed reading into the **Analysis** the 6 931 records. In the rest of this exercise, you will conduct simple exploratory analysis according to the data-cleaning plan to find possible remaining errors.

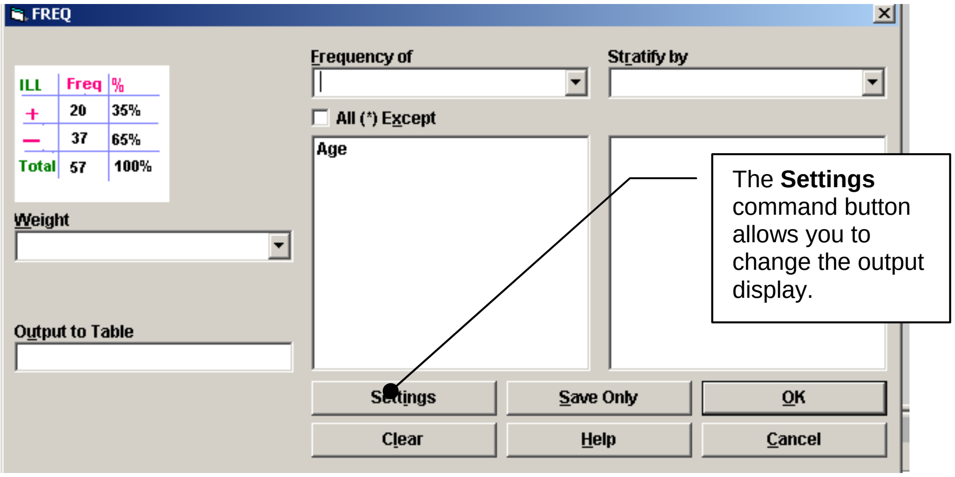
Obtaining a Frequency

According to our data-cleaning plan, we want to review age to ensure that all women included meet the age eligibility criteria. We are also specifically interested in those limited instances of women aged 12 who are pregnant, since these data will be carefully scrutinised by program planners. To calculate a frequency of *Age*:

Steps to calculate age frequency

1. Under the Statistics folder on the Command Tree, click the **Frequencies** command.

2. Select *Age* from the **Frequency of** list box.



The **Settings** command button allows you to change the output display.

3. Click the **Settings** button to change statistics to **None**. Check the **Include Missing** to ensure that if any ages were mistakenly not entered, we would see these.

4. Click **OK**.

5. Click **OK** in the FREQ box.

6. Review the results in the **Analysis Output** window. Note that there are two records where *Age*=12. These records must be manually reviewed according to our data-cleaning plan. To do that, we need to know the unique *pt_key* numbers that correspond to these two records to identify the correct data collection form and the correct record number.

Using Analysis to Find Specific Records

There are many ways to identify specific forms or electronic records based on a value in the database. For example, we could manually review all of the data collection forms to see which ones list an age of 12. Conversely, we could search each electronic record in the Enter application in Epi Info. This would take some time, however, if the number of records is large.

Using the Find and Select commands to search the database

Instead of searching manually through the database, we can also use the computer to search for us, as we saw in Epi Info's **Enter Data** tool. For example, we used the **Find** command in **Enter**. Similarly, in **Analysis**, we can use the **Select** command to locate a specific record.

In **Analysis**, to identify those records and the *pt_key*, we want to select those records where age <13, then list the records, either with *pt_key* only, or with all fields.

Selecting a Sub-Set of Records

1. Under the **Select/If** folder in the Command Tree, click the **Select** command and type the expression age<13 in the Select Criteria box.

2. Click **OK.**
   You should see two records in the current data set.

Current View: *C:\ANC_Suri\ANC2002\sys02c.mdb:ANCSurveillance2*
Select: *(Age < 13)*
Record Count: *2 (Deleted records excluded)*
Date: *8/01/2003 12:59:54 PM*

Obtaining a Line Listing of a Sub-set of Records

1. Under the **Statistics** folder on the Command Tree, click the **List** command to create a line listing of the two records.

2. Click **OK**.

Making changes to the data

*Epi Info can display line lists as an HTML table in a Grid spreadsheet. If you select **Allow Updates**, you can make changes to the data. However, changes to the database are permanent and no record of the change will be kept electronically.*

Selecting variables to display

The asterisk (*) represents all variables available in the database. To list only selected variables, replace the asterisk with the name of the fields in the Variables list. Note that you can also display "All Except" the listed variables by selecting this option.

Try it yourself!

Activity 1, Use Original Forms to Find Errors

Find the original data-entry forms for the two records with Age < 13 in "Appendix I – Selected original data-entry forms for 2002 sites" to compare *Age* in the database with the printed age on the form.

If an error exists, fill out your data-entry audit log in Appendix H for the report in question. Complete the audit log except for the method of resolution. In the next section, you will identify methods for changing data in the data set that will allow you to correct the errors documented in the audit log.

Canceling the Select Criteria

Select statements remains active until the user cancels them or a new file is read.

Multiple select is the same as issuing selects with a conditional AND statement. For example, age<13 AND Pt_key="511133" will return only the record(s) that meet both conditions where age<13 AND Pt_key="511133"

1. Click on **Cancel Select** to remove the select criteria.

2. Click **OK**.

Try it yourself!

Activity 2, Complete Data Analysis Plan

Complete the rest of the data analysis plan for the 2002 data, beginning with the additional frequencies and tables. At minimum, you should review the following:

- frequency of site to ensure minimum sample size has been achieved
- gravidity such that at least one pregnancy is listed
- frequency of syphilis results
- frequency of HIV results for missing values
- table analyses to check for consistency of parity and gravidity, and key outcome variables (*HIV_res*) by site
- consistency of dates.

Note that **Tables** analyses, similar to the **Frequency** outputs, are used when you want to cross-tabulate frequencies of multiple variables. To use **Tables** in Epi Info, you must select an exposure variable (X variable) and an outcome variable (Y variable).

In the case of gravidity and parity checks, *Par* is the exposure variable (column heading) and *Grav* is the outcome variable (row heading). For most 2x2 Tables, you can also use the frequency command by stratifying on the exposure variable in the **Stratify Dialogue Box**.

Identify any inconsistencies and note them in your data audit log as you did with age. If you find an anomaly in the data, rather than a problem with a specific record, write the problem on a single line of the data-entry audit log and talk with the consultants at the end of the exercise about how to resolve this issue during data cleaning.

Activity 3, Review Program Code

Review the program code in the Program Editor window. The steps that you took should be listed there. Note how Epi Info places the commands in capital letters and the variables in lowercase. Place your cursor in the Program Editor window and click to activate the window. Document your program code to show that you are 1) reading the ANC 2002 database and 2) identifying those records that have age<13.

*Use * to begin comment lines in the Program Editor. Epi Info will ignore lines beginning with * when processing program analysis code.*

Exercise 7
Data Cleaning

Overview

What this exercise is about

In Exercise 6, we identified an error of a client's age after reviewing the age frequencies and the original data forms. We also identified other possible data-entry problems, such as the high HIV prevalence due to sample degradation at Site 17. In the audit log, the group decided how to resolve these errors and made a note of the resolutions.

In Exercise 7, we will create a clean dataset for the 2002 records by editing erroneous data values and outputting a new data table containing no known errors. Once you have completed cleaning the 2002 dataset, you will follow the same data-cleaning plan for the Epi Info 2001 ANC dataset (*C:\ANC_Suri\ANC2001\sys01.mdb:ANCSurveillance*) containing 6 762 records.

Recall that the data from 2001 was only cursorily analysed to determine HIV prevalence; the Surveillance Team did not conduct in-depth exploratory and statistical analyses. Based on the results of your data-cleaning exercise, the team will make notes in the 2001 audit log, resolve differences and edit the 2001 dataset in preparation for more extensive analysis of trends.

What you will learn

At the end of the exercise, you will be able to:

- list the benefits and limitations of using Enter, Analysis, and Visualize Data to fix simple data-entry errors.
- use If/Then and Assign statements to replace values in a cleaned data set.
- use recodes to standardise responses for a text value statement.

Starting location

Enter Data, C:\ANC_Suri\ANC2002\sys02c.mdb: ANCSurveillance2

Resources

Appendix H – HIV Surveillance Data Audit Log
Appendix I – Selected original data-entry forms for 2002 sites

Editing Data Values

Epi Info allows you to edit data using a variety of tools within the application. Each method has benefits and drawbacks to its use. You may be familiar with Methods 1 through 3 in the table below, and may already be able to list some of the benefits and drawbacks. Method 4, Visualize Data, is a tool for editing data that has not previously been mentioned.

Methods for Editing Data

Method	Benefits	Drawbacks
1. Enter Data **Location:** ▪ In Main Menu ▪ Click Programs ▪ Select Enter Data	▪ Easy to edit data by using the original data-entry screen ▪ Unlikely to select wrong value for coded responses ▪ Records cannot be mistakenly deleted ▪ Deleted records are visible but excluded during analysis ▪ Allows check code to be run on values as they are entered	▪ Directly changes value in original data set with no record other than the audit log, except for deletion of records
2. Data Compare **Location:** ▪ In Main Menu ▪ Click Utilities ▪ Select Data Compare	▪ Directly changes value in the dataset ▪ Records cannot be mistakenly deleted ▪ Changes made to the database can be “undone” by re-editing the original database	▪ Directly changes value in original data set with no record other than the audit log ▪ Only variables that differ between the main and double data-entry datasets are highlighted ▪ Only highlighted responses can be changed ▪ Check code is not run on data entered in this application
3. Analyze Data **Location:** ▪ In Main Menu ▪ Click Programs ▪ Select Analyze Data	▪ Records cannot be mistakenly deleted ▪ Provides a program file that can be saved to document changes to individual records ▪ Complements data audit log documentation	▪ Program files must be saved to reproduce the steps for data cleaning ▪ Program files require more time and documentation than simply changing a value ▪ Check code is not run on data entered in this application ▪ Using the list/update command allows changes outside ENTER
4. Visualize Data **Location:** ▪ In Main Menu ▪ Click Utilities ▪ Select Visualize Data	▪ Allows for direct editing in the original dataset ▪ Allows records to be permanently deleted	▪ Complex; difficult to understand and use ▪ Directly changes value in original data set; no record other than the audit log ▪ Check code is not run on data entered in this application

Editing Data Values, continued

Regardless of the method selected, it is critical to ensure that data are not unintentionally changed or corrupted during the editing process. Therefore, you should always make a backup copy of your database before edits are made.

Using Enter Data and Analyze Data commands to edit data

For the purposes of Exercise 7, we will use Method 1 **Enter Data** to delete the record of the non-pregnant woman whose *Pt_key*="511065". To edit the age value and to exclude records belonging to Site 17, we will use **Analyze Data** (Method 3).

Once the edits have been made, a new cleaned file will be available in preparation for analysing our data while also providing us with documentation of changes we made to the data.

Deleting Records in Epi Info

Steps to delete records

1. From the File menu, click **Enter Data**.

2. Open *C:\ANC_Suri\ANC2002\sys02c.mdb* project and the *ANCSurveillance2* table.

3. Using **Find**, search for records where *grav* = 0.

4. Double-click on the grey box to the left of the record where *pt_key* = "511065" to bring the record back to the data-entry view.

5. Click on the **Mark Record as Deleted** button in the left panel. The record status has been changed for the record of the non-pregnant woman. Also, notice that the entry boxes are no longer editable.

6. Exit the Enter program.

*Deleted records are still displayed in **Enter** in a disabled format and have a red 'Deleted Record' flag above the record number in the bottom left of the command tree. Data cannot be edited in **Enter** and can be excluded from analysis when a record is marked as deleted. A record can be undeleted by clicking on the **Undelete** button.*

Using If/Then and Assign Statements in Analysis to Replace Values

Steps to replace errors

If/Then statements allow changes to a variable if a certain condition is met. You should be familiar with If/Then statements from their use in creating check code in **Enter** during Exercise 3. This time, in **Analysis**, we will use the If/Then statement to replace an error in the database. For example, if *Pt_key*= “511133” then *Age* should be 21 instead of 12.

1. From the **File** menu, click **Analyze Data**.

2. **Read (Import)** *C:\ANC_Suri\ANC2002\sys02c.mdb* from **Analysis**. Ensure that the Project prompt reads: *C:\ANC_Suri\ANC2002\sys02c.mdb*.

3. Select the table *ANCSurveillance2* from the command tree.

Current View: *C:\ANC_Suri\ANC2002\sys02c.mdb: ANCSurveillance2*
Record Count: *6930 (Deleted records excluded)*
Date: *10/01/2003 9:10:40 AM*

Note that the deleted records in Epi Info Analysis are not included, unless you specify in the Options Set folder to also process these records

4. Click **IF** in the Command Tree under **Select/IF.**

Steps to replace errors, continued

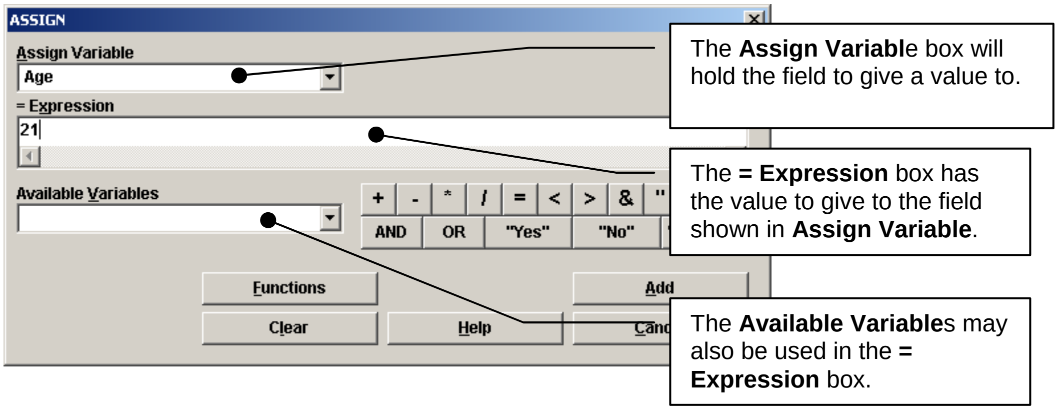
5. Select *Pt_key* from the **Available Variables**.
6. Type = “511133” to indicate that only that record should be selected. Include the quotation marks. The *Pt_key* field is text.
7. Select **AND**.
8. Choose **Age.**
9. Type =*12*

   Although technically unnecessary, the conditional **AND** statement will ensure that you did not misidentify the patient record, since both statements must be true to continue. Because *pt_key* is a text variable, we include it in quotations. Conversely, because age is a numeric variable, we do not put quotations around the number.

Steps to replace errors, continued

10. Click **THEN**.

11. Choose **Assign** from the 'Then Block' tree structure under the Variables commands. A pop-up window will appear labeled **ASSIGN**.



Assign Variable: Age

= Expression: 21

The **Assign Variable** box will hold the field to give a value to.

The **= Expression** box has the value to give to the field shown in **Assign Variable**.

The **Available Variables** may also be used in the **= Expression** box.

12. Select **Age**.

13. In the **= Expression** box, type *21*.

14. Click the **Add** button.

15. Click **OK.**

The following commands will be visible in the program editor:

```
READ 'C:\ANC_Suri\ANC2002\sys02c.mdb':ANCSurveillance2
IF Pt_key=“511133” AND Age=12 THEN
    ASSIGN Age=21
END
```

To denote ‘does not equal’ in your condition, use the combination of the Less Than and the Greater Than signs (<>). For example, Age does not equal 12 is expressed as Age <>12.

Try it yourself!

Activity 1, Use IF/THEN Statement to Clean Data

a. Using an IF/THEN statement, correct the *grav* value for *pt_key* = "511173"

b. Site 17 has been excluded from analysis because of laboratory testing problems. Use the **Select** command to identify records belonging to Site 17 and select sites not equal to "17." Be sure to update your data audit log.

c. Review the changes to the fields for *sit_num*, *age,* and *grav* to ensure that the changes were accepted. Place a check mark next to each statement after you have verified the changes. **Please do not proceed until each item is checked.**

□ There is only **one** record with an age of 12.

□ There is no record with the gravidity = 0.

□ Site 17 is excluded from the data set, giving a new total of **6 604 records**.

If you have correctly verified the changes to the data set, please proceed with the rest of the exercise. Otherwise, repeat any instructions that have not produced results to satisfy the checklist. Ask for assistance if necessary.

Saving Changes to the Data File Using WRITE

All edits and updates that you make to your data during the Analysis session are done in a copy of the original data. **If you exit without saving the changes by using the Write/Export command, they will be lost.** Make sure that you save the new data set with a new table name, or your original data will be overwritten!

Steps for saving changes

1. Click **Write/Export** on the Analysis command tree to bring up the pop-up box.

2. Ensure that the **ALL** radio button is checked or select all the variables for output.

3. Under Output, choose **Replace**.

   Since the *cleaned02* data table is new, this selection does not matter. However, in the future, it will be important to **replace** the data file rather than to **append**. Appending data results in multiple copies of the data file in the data table.

4. In the **File** name, either click the **...** button and choose *sys02c.mdb* or type *C:\ANC_Suri\ANC2002\sys02c.mdb* into the prompt.

5. In the Data Table name, type *cleaned02*.

6. Click **OK.**

The following text should appear in the Program Editor:

```
WRITE REPLACE “Epi 2000”
'C:\ANC_Suri\ANC2002\sys02c.mdb':cleaned02 *
```

Your new data table *cleaned02* with the changes will be created for future use.

Saving Program Files

Saving program text file for future use

The program text file you created to clean the data file can be used in the future to clean other datasets and to document your activities. To save the program file:

1. Click **Save**.

2. Ensure that the **Project File** is *C:\ANC_Suri\ANC2002\sys02c.mdb*.

3. Type *anc2002clean* into the **Program** prompt.

4. Click **OK**.

This will save a copy of the program inside the project (MDB) file. This program can then be retrieved and run at a later time rather than rewriting and reconstructing the program. This program also serves as important documentation of the changes that you made to the original data.

Try it yourself!

Activity 2, Prepare 2001 Data-Cleaning Plan

Similar to your data-cleaning plan for the 2002 dataset, you should also have a data-cleaning plan for the 2001 data set. As noted in the case study, the data set for 2001 was double-data entered; however, the data were not cleaned before the analysis of results.

Using the same or similar practices used for the 2002 plan, you should be able to quickly identify the steps that will be required to produce a clean data file for 2001; they may even be the same steps. Create a written data-cleaning plan for the 2001 database.

Activity 3, Begin Analysis of 2001 Dataset

Click **New** in the **Program Editor**. We will have a new program for the steps used to clean the 2001 data set.

Following the steps in your data-cleaning plan, use **Analysis** to read the 2001data set. To get you started, the 2001 data set is located in *C:\ANC_Suri\ANC2001* and can be accessed using Epi Info **Analysis**.

Begin by using the **Read (Import)** command in the left command tree. Click **Change Project** and navigate to the *C:\ANC_Suri\ANC2001\Sys01.mdb project*.

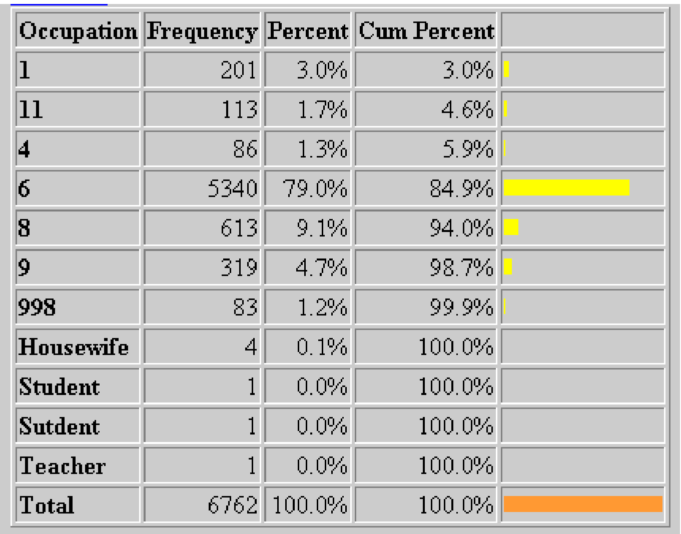
Select the *ANCSurveillance* table. Once it is selected, you should have 6 762 records. You can now begin your exploratory analysis of the 2001 data file.

Note any anomalies for the dataset in your data-entry audit log. Use the Data-Cleaning Audit Log at the end of Appendix H, page H.2-1 to record any changes to be made. Do not make any changes at this time. We will be instructed on how to make the changes to the 2001 data set in the rest of the exercise.

Recoding Text Fields for Editing Purposes

From Activity 3, you may have noted that the variable *occup* contains inconsistent text entries and a problem with miscoded unknown and missing values.

The frequency of *occup* is shown:



Occupation	Frequency	Percent	Cum Percent	
1	201	3.0%	3.0%	
11	113	1.7%	4.6%	
4	86	1.3%	5.9%	
6	5340	79.0%	84.9%	
8	613	9.1%	94.0%	
9	319	4.7%	98.7%	
998	83	1.2%	99.9%	
Housewife	4	0.1%	100.0%	
Student	1	0.0%	100.0%	
Sutdent	1	0.0%	100.0%	
Teacher	1	0.0%	100.0%	
Total	6762	100.0%	100.0%	

Steps to recode values

To correct these values using the **Recode** statement, follow the steps below:

1. Select **Define** from the analysis tree menu.
2. Create a new variable called *occup1*.
3. Select **Recode**.
4. Select *occup* as the **FROM** variable and *occup1* as the **TO** variable.

The first column (value) holds the original values. The second (middle) column is ignored because we are recoding a text variable. The third column (recoded value) will hold the new value.

Text must be enclosed in double quotation marks. The word ELSE may be used to indicate all values not falling in the specified ranges. Recodes take place in the order stated; if two ranges overlap, the first in the list will apply.

5. Using the 11 values displayed in the frequency output, recode the responses to the new values below:

Value		
(blank =other)	**To Value (if any)**	**Recoded Value**
"1"		"1"
"4"		"4"
"6"		"6"
"8"		"8"
"9"		"9"
"11"		"11"
"998"		"998"
"Student"		"4"
"Sutdent"		"4"
"Housewife"		"6"
"Teacher"		"9"

*Any value not included in the **Recode** will be changed to NULL or missing in the data table. Include the unchanged values in recode (i.e., recode "1" to "1" and "2" to "2").*

6. Click **OK** to exit the **Recode** dialog box once all of the values have been entered.

7. Create a frequency of the table of the *occup1* variable to ensure that the values have been correctly recoded.

8. Compare the new frequency to the previously created frequency to ensure that all values were recoded properly. Use the **Set** command at the bottom of the command tree to check **Include Missing** so there is a true account of all the values represented, even those with no value entered.

Updating data-entry audit log

As was done to the 2002 data set, be sure to update your data-entry audit log to document the changed values and save the changes to the 2001 data set. To save the changes, use the **Write (Export)** command. Steps are listed below to ensure that you have saved the changes correctly.

Saving the Changes

Steps for saving changes

To save the new values for subsequent analysis, use the **Write (Export)** command.

1. Select the **Write (Export)** command from the Analysis menu tree.

2. Select **Replace** under **Output Mode**.

3. Select *C:\ANC_Suri\ANC2001\sys01.mdb* by clicking on the **…** button under **File Name**.

4. In the **Data Table** prompt, create a new table by typing *cleaned01* in the data-entry box.

5. Click **OK**.

6. Click **Read (Import)**.

7. To verify that the table *cleaned01* was saved properly, click the **All** radio button to see the new table. (This is a data table, not a view.)

8. Under the **Variables** folder, click **Display**.

9. Click **OK**. A data dictionary of the new *cleaned01* table will be shown.

Don't forget to:

- document the program code in the Program editor
- save your program code in the Program Editor to use for future data cleaning activities.

Exercise 8
Preparing Data for Analysis

Overview

What this exercise is about

Before analysing data, it is useful to have a plan that describes the types of analyses to be done. During this exercise, we will have the opportunity to create a data-analysis plan and construct a cleaned data file using data from the 2000, 2001 and 2002 ANC rounds. To do this, 2000 ANC data from Epi Info DOS and 2001 data from Epi Info will be appended to the 2002 ANC data set.

At the end of this exercise you should have a single data file in Epi Info that contains:

- 5 230 records from 2000
- 6 762 records from 2001
- 6 604 records from 2002

In addition, new variables will be created for the recoding of *Age* to *AgeGroup* and *Visit Date* to *Year*, for the year to which the survey data belong. Labels will be added using the recode function to simplify presentation during analysis. Missing values will be consistently coded to facilitate analysis.

What you will learn

At the end of the exercise you will be able to:

- develop a data-analysis plan
- understand how to open, read and write Epi 6 file formats in Epi Info
- append cleaned databases into a single Epi Info file for analysis
- create recoded variables useful for analysis.

Starting location

Epi Info Main Menu, *C:\ANC_Suri\ANC2002\sys02c.mdb::cleaned02*

Developing a Data-Analysis Plan

When the ANC survey was initially designed in Suri, stakeholders identified the basic data elements required to describe HIV prevalence among the population sample. These basic data needs influenced the design of the data collection form and the manner in which data were collected.

Developing a data-analysis plan prior to beginning analysis helps data users to further think analytically about how they will describe their results. In ANC sentinel surveillance programs, analyses typically include tables of general population characteristics, some comparisons of HIV prevalence among various sites or among specific sub-groups of the population and comparison of trends over time if data are available.

Statistical analyses to be used

For analysis purposes, the following data-analysis plan will be adhered to:

- **Univariate analysis** – simple descriptive statistics of the sample population's demographic characteristics, including the HIV and RPR prevalence for 2002 according to:
  - site
  - district
  - age group
  - marital status
  - educational level
  - residence
  - gravida
  - parity
  - occupation.
- **Bivariate analysis** – crude comparison of HIV prevalence between urban and rural residents and between younger women (<25) as compared to older women (>=25) for 2002. Age-standardised comparison of HIV prevalence in urban and rural women.
- **Multivariate analysis** – Comparison of HIV prevalence in clinics over time using 2000, 2001 and 2002 data

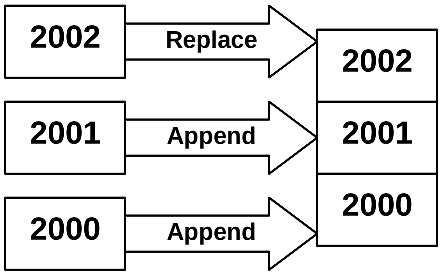
In addition to these statistical analyses, graphs and bar charts can provide information so that it is easier to understand, and will be generated to illustrate key results.

Creating a single data file

To prepare a file for data analysis, a single file that will include data from the three years of data collection (2000, 2001 and 2002) will be created. Special summary variables, *AgeGroup* and *Year*, will be added to aid in analysis. Text values will be recoded to create labels for the tables, graphs and maps. Missing values will be recoded into a format recognised as "missing" by Epi Info, which will allow us to include or exclude missing values more easily in our analyses.

By creating a single data set that contains clean data for the three years of interest, we will also be able to perform trend analysis of HIV prevalence later.

Create a Single File from All Three Years



Creating an Epi Info Data Analysis File Using Two Epi Info Databases

When undertaking new tasks, you should start a new program in the Program Editor.

Steps to create a project with three years of data

To create the ANCAll project containing the three years of data, first:

1. Click **New** in the **Program Editor** to start a new program.

2. Click **Read (Import).**

3. Click the **Change Project** command button to select *C:\ANC_Suri\ANC2002\sys02c.mdb* if it is not already selected.

4. Select the Show: **All** radio button.

5. Select the *Cleaned02* table.

6. Click **OK**.

To export the 6 604 records in *cleaned02* data table to a new project that will contain all three years of data in a single table, follow the steps below:

7. Select **Write (Export)** from the command tree.

8. Verify that the Output Mode is **Replace.**

*When you initially create the data table, either **Append** or **Replace** can be selected. Append will add records to the existing data table (which is currently empty) while Replace deletes all records in the existing data table.*

9. Verify that the output format is Epi 2000.

10. Click on the three dots to the right of the **File Name** box and navigate to the *C:\ANC_Suri\Analysis\* folder. Type *ANCall* as the file name. Click **Save**.

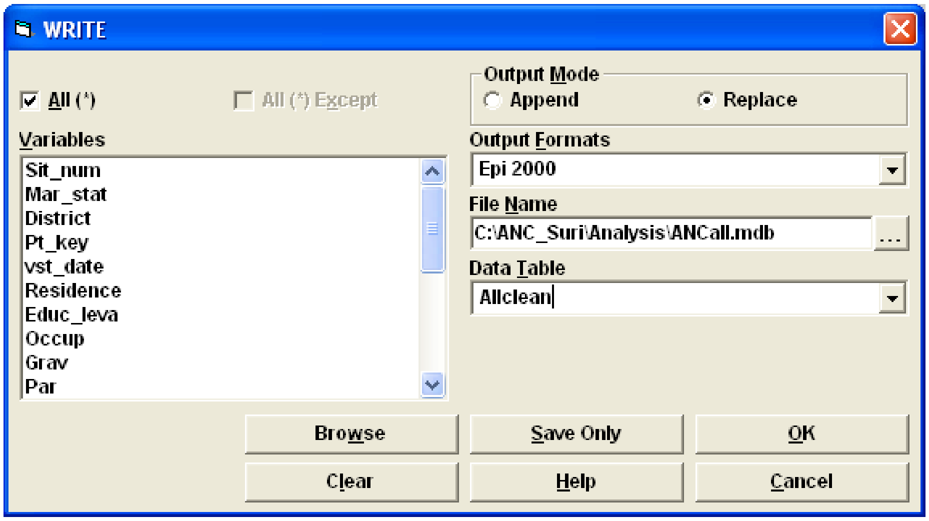
Steps to create a project with three years of data, continued

This will create a new project called *ANCall.mdb*, which will contain the 2000, 2001 and 2002 data. This file will be located in the Analysis folder.

11. Type *Allclean* as the Data Table name into which the 2002 data will be saved.

12. Click **OK**.

For reference, the **Write** box at Step 11 should appear as follows:



WRITE

☑ All (*) □ All (*) Except

Variables: Sit_num, Mar_stat, District, Pt_key, vst_date, Residence, Educ_leva, Occup, Grav, Par

Output Mode: ○ Append ● Replace

Output Formats: Epi 2000

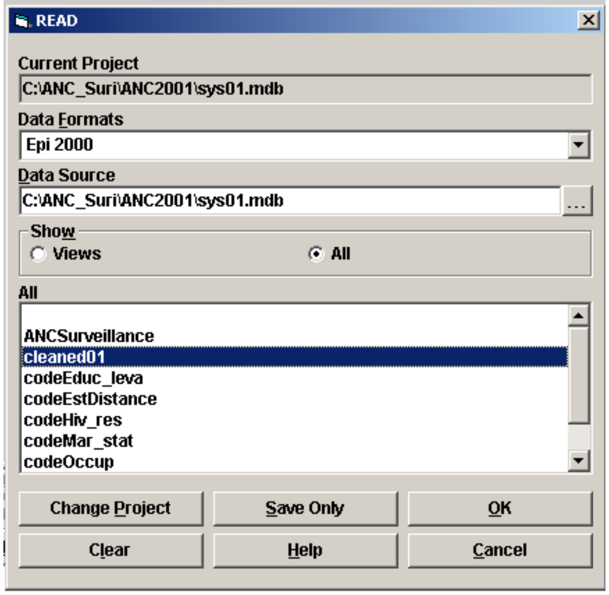
File Name: C:\ANC_Suri\Analysis\ANCall.mdb

Data Table: Allclean

Try it yourself!

Activity 1, Append 2001 Data

To append the 2001 data to *C:\ANC_Suri\Analysis\ANCall.mdb* to the *Allclean* data table, repeat steps 2-6 using the *C:\ANC_Suri\ANC2001\sys01.mdb:cleaned01* project and *cleaned01* data table. Note that you will append, not replace, the data set for 2001. At the completion of step 4, the **Read** pop-up box should appear as below:


READ

Current Project: C:\ANC_Suri\ANC2001\sys01.mdb

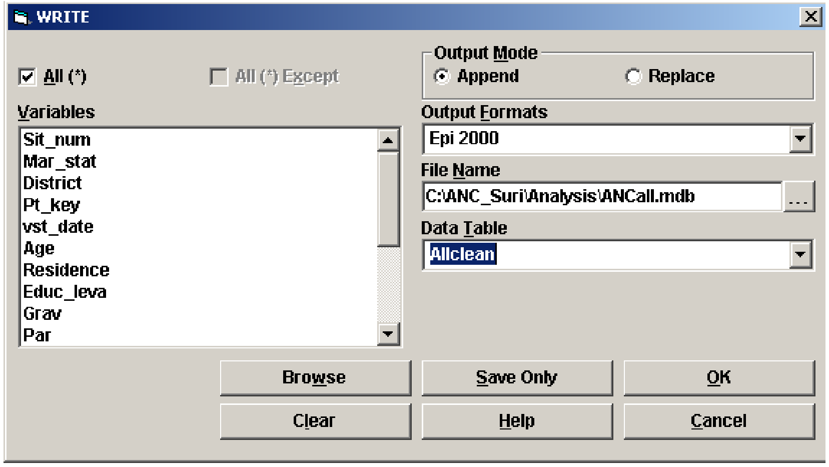
Data Formats: Epi 2000

Data Source: C:\ANC_Suri\ANC2001\sys01.mdb

Show: ○ Views ● All

All:
- ANCSurveillance
- cleaned01
- codeEduc_leva
- codeEstDistance
- codeHiv_res
- codeMar_stat
- codeOccup

Continue with steps 7 through 12, but use 2001 data and **Append** these data in step 7; 6 762 records will be appended to *ANCAll.mdb* project. For reference, the **Write** box at Step 11 should appear as follows:


WRITE

☑ All (*)   □ All (*) Except

Variables:
- Sit_num
- Mar_stat
- District
- Pt_key
- vst_date
- Age
- Residence
- Educ_leva
- Grav
- Par

Output Mode: ● Append ○ Replace

Output Formats: Epi 2000

File Name: C:\ANC_Suri\Analysis\ANCall.mdb

Data Table: Allclean

Activity 1, Append 2001 Data, continued

To ensure that you have appended the data tables correctly, **Read** the *Allclean* table in the *ANCall* database. A total of 13 366 records should now be shown in the data table.

Appending Data from an Epi Info 6 (DOS) Format

Steps to read an Epi Info 6 DOS file

While the 2001 and 2002 data were in an Epi Info data format, the 2000 data were entered into Epi Info 6 DOS. To read an Epi Info 6 DOS file, follow the steps below:

1. Click **Read (Import).**

2. Select **Epi6** as the file format from the **Data Format** drop-down box.

3. Click on the **...** button to the right of the **Data Source** text box.

4. Navigate to *C:\ANC_Suri\ANC2000*.

5. Select the *anc2000.rec* file containing the year 2000 data.

6. Click **OK**.

The following text should appear in the **Program Editor** window:

Current View: *C:\ANC_Suri\ANC2000\anc2000.rec*
Record Count: *5230 (Deleted records excluded)*
Date: *10/01/2003 9:10:40 AM*

7. Click **Write (Export)** in the tree command box.

   a. Verify that the output mode is append.
   b. Verify that the output format is Epi 2000.

8. Navigate to *C:\ANC_Suri\Analysis\ANCall.mdb* in the project file name prompt.

Steps to read an Epi Info 6 DOS file, continued

9. Type or select *Allclean* from the drop-down prompt as the **Table Name** into which the 2000 data will be saved.

10. Click **OK**.

11. Save the program code in the Program Editor.

Importing the .REC file into a new table will automatically generate a view for that data table.

To verify that all 18 596 records are in the data table called *Allclean*, **Read** the *C:\ANC_Suri\Analysis\ANCall.mdb: Allclean* data table.

Modifying Data for Data Analysis

In Exercise 7, we cleaned data using IF/THEN or RECODE statements. These statements are also valuable for creating new variables or modifying variables to make our analyses easier to understand. For example, we will need to recode the text or number codes that we had used to indicate "missing" or "unknown" during data-entry to a value that Epi Info recognises as missing. We will also group certain numeric fields, which have many possible responses, into a smaller number of categories to simplify data presentation. Finally, we will create labels for our variables and create new variables.

In Exercise 7, we learned that with the RECODE statement, all values for a variable, even the unchanged values, must be included; otherwise, values left out will be missing in the recoded variable. Because of this, the RECODE statement is most useful when creating a new variable or recoding all values of a variable. If just recoding certain values, it is often easier to use an IF/THEN statement. We will see examples of both of these approaches in the following exercises.

Recoding Missing Values to a Value Recognised By Epi Info as Missing

When we created the ANC database, we created codes to indicate missing or unknown responses. These codes (for example 998 or 999) are not recognised as missing values by Epi Info and, therefore, cannot be easily included or excluded from analysis using options available in the **Analysis** window. Although knowing which responses were missing versus unknown may be important for survey quality assurance, we will combine missing and unknown values into a general category of missing values for our analyses.

Steps to recode missing or unknown numeric data

To recode the values that we used to indicate missing or unknown (998 or 999) for the numeric variable *Par* to the Epi Info code for missing, follow the steps below:

1. Click **New** in the **Program Editor** to create new program code.
2. **Read** the *Allclean* data table from *C:\ANC_Suri\Analysis\ANCall.mdb*.
3. Click **IF** in the Command Tree under **Select/IF**.
4. Select *Par* from the **Available Variables**.
5. Type *>=998* to indicate that records with values of 998 or 999 should be selected. Note that because *Par* is a numeric variable, we do not use quotation marks.
6. Click **THEN.**
7. Choose **Assign** from the ‘Then Block’ tree structure under the Variables commands.
8. Select *Par*.
9. In the =Expression box, either type =(.) or select the “Missing” value from the choices found under the =Expression box.
10. Click the **Add** button.
11. Click **OK**.

Steps to recode missing or unknown numeric data, continued

The following commands will be visible in the program editor:

```
READ 'C:\ANC_Suri\Analysis\ANCall.mdb':Allclean
IF Par>=998 THEN
    ASSIGN Par= (.)
END
```

Try it yourself!

Activity 2, Recode the Missing/Unknown Values for the Gravidity Variable

Recode the missing or unknown values for the gravidity variable to the code that Epi Info recognises as missing. Because we are recoding only a few of the possible values for the *Grav* variable, we will use an IF/THEN statement. Refer to the *Par* example above to guide you if necessary.

Recoding Numeric Fields for Data Analysis

Numeric fields, which have many possible responses, are usually grouped into a smaller number of categories for data analysis. For example, descriptive analyses of HIV data typically use five-year age group intervals, as recommended by WHO. To recode the numeric values of *Age* to a text variable *AgeGroup*, follow the steps below:

Steps to recode a numeric value to a text value

1. **Define** a **Standard** variable called *AgeGroup*.

Standard variables created with the Define command persist only for the table for which they are created. If you read a new database, all defined standard variables will be lost. To make the variable permanent, before reading a new table or project you must write out the table, using the ***Write (Export)*** *command.*

Global variables retain values across tables in databases for as long as the Epi Info program that defined the global variable is open.

Permanent variables hold single values only and can be saved as a part of Epi Info system file. The variable is available to any Epi Info database.

Steps to recode a numeric value to a text value, continued

2. **Recode** *Age* to the new variable, *AgeGroup*.

Numeric recoded ranges are separated by a space, hyphen, and space, as in 1 – 5. Negative values are permitted, as in -10, -9 and -8. Note that *AgeGroup* is a character variable and therefore requires quotes around the values.

Value (blank = other)	To Value (if any)	Recoded Value
12	14	"12 – 14"
15	19	"15 – 19"
20	24	"20 – 24"
25	29	"25 – 29"
30	34	"30 – 34"
35	39	"35 – 39"
40	44	"40 – 44"
45	49	"45 – 49"

The words LOVALUE and HIVALUE may be used to indicate the smallest and largest values represented in the database, respectively.

3. Click **OK** when finished. The Recode statement will appear in the Program Editor.

4. **Write (Export),** selecting the **Replace** output method to the *Allclean* table.

5. **Read (Import)** *Allclean* table.

6. Verify recode of *AgeGroup* using **Frequency**.

Epi Info sometimes has problems maintaining recodes in its temporary memory. When this happens, you will receive an error notification requiring you to exit Epi Info, which means you will lose all of the work that you had done recoding. It is good general practice to write and re-read the file after every few recodes to minimise these types of Epi Info errors.

You can also recode variables by creating an output table, making new text values to replace the numeric values for the field, and then using the **Relate** command to relate the output table to the main table, incorporating the new values in the process. We will try this using the variable *Occup.*

Steps to recode a numeric value to a text value using the relate command

1. Click on the **Frequency** command. In the **Frequency of** box, select *Occup*. In the box in the bottom left-hand corner where it says **Output to Table**, type in **Occup1.**

2. Click **OK.**

3. Read in the new table by selecting show **All** views and then selecting **Occup1.** You should have a record count of 8.

4. Run a **List** of all variables (*****), making sure that you select **Allow Updates** under display mode. The following table should come up.

Occup	VARNAME	COUNT
1	Occup	341
10	Occup	588
11	Occup	191
4	Occup	237
6	Occup	14459
8	Occup	1058
9	Occup	1156
998	Occup	566

5. In the middle column under the heading VARNAME, type in the new values in the corresponding row.

Occup	VARNAME	COUNT
1	"01 – Business"	341
10	"10 – Not employed"	588
11	"11 – Other"	191
4	"4 – Student"	237
6	"6 – Housewife"	14459
8	"8 – Laborer"	1058
9	"9 – Professional"	1156
998	(.)	566

6. Close the window.

7. **Read** the *Allclean* table.

Steps to recode a numeric value to a text value using the relate command, continued

8. Click on the **Relate** command, which is directly under the **Read** command. Select the Show **All** button if it is not already selected. Now select the *Occup1* table from the menu. Click on **Build Key.**

9. The **Relate – Build Key** box will pop up. Select *Occup* from the **Available Variables** drop-down menu. Now click on **Related Table.** You will see *Occup* drop into the box underneath Current Table(s).

10. Select *Occup* from the **Available Variables** drop-down menu. Click **OK.**

11. Click **OK** again.

12. Under Key, it should say Occup :: Occup. Click **OK.**

13. Click on the **Define** command. Call the new variable *Occupation.* Click **OK.**

14. Click on **Assign.** In the **Assign Variable** box, type or select *Occupation.* In the =Expression box, type or select the variable *VARNAME.* Click **OK.**

15. Run a **Frequency** of *Occupation* to make sure the recode was successful.

Recoding Text for Data Analysis

The name of clinic sites is frequently needed when conducting analyses. For the purposes of creating the unique patient ID on the data collection form, the variables *sit_num* and *District* were stored as text but with numeric values to simplify data entry. For analysis purposes, however, we want to show the site and district names.

Try it yourself!

Activity 3, Recode the District Variable

Recode the district variable. Use the *occup* example in Exercise 7 to guide you if necessary. Each time, define the new variable, then recode. Remember to save your work. District values should be as follows:

Activity 3, Recode the District Variable, continued

District	District1
"1"	"Tibul"
"2"	"Mandor"
"3"	"Rikura"
"4"	"Yemenia"
"5"	"Insa"
"6"	"Karafam"
"7"	"Ashra"

Recode variables *Residence, Educ_leva,* and *Mar_stat* using the information below. Note that we can recode missing values to the Epi Info "Missing" code at the same time that we are recoding our text variables.

Educ_leva	Education1
"1"	"1 -None"
"2"	"2 – Primary"
"3"	"3 – Secondary"
"4"	"4 – Higher"
"98"	(.)

Residence	Residence1
"1"	"1 – Urban"
"2"	"2 – Rural"
"98"	(.)

Mar_stat	MarStatus1
"1"	"1 – Single"
"2"	"2 – Married"
"3"	"3 – Divorced"
"4"	"4 – Widowed"
"98"	(.)

Recode *HIV_res* to HIV, *RPR_res* to RPR and *TPHA_res* to TPHA, using the following information:

Activity 3, Recode the District Variable, continued

HIV_res	HIV
"1"	"1 – Positive"
"2"	"2 – Negative"
"98"	(.)

RPR_res	RPR
"1"	"1 – Positive"
"2"	"2 – Negative"
"98"	(.)

TPHA_res	TPHA
"1"	"1 – Positive"
"2"	"2 – Negative"
"98"	(.)

Recoding Text for Data Analysis with More Than 12 Responses

Recoding from one value to another as was done in Exercise 7 for the *occup* variable or above for the *district* variable was simple when 12 or fewer responses required recoding. In Epi Info, however, recoding for variables that have more than 12 responses can be challenging. The example below is provided to create a *SiteName* variable, based on *Sit_num*, which will be used during analysis.

Steps for recoding >12 responses

1. **Define** *SiteName* from the Analysis menu tree.
   SiteName will save the recoded text values for sites 01 – 10.

2. **Define** *tempSite1* from the Analysis menu tree.
   tempSite1 will save the recoded text values for sites 11 – 19.

3. Select the **Recode** command from the Analysis menu tree.

4. Select *Sit_num* as the **From** variable and *SiteName* as the **To** variable.

5. Using the following information, recode *SiteName*:

Sit_num	SiteName
"01"	"Banket"
"02"	"Chema"
"03"	"Chickry"
"04"	"Cholai"
"05"	"Danu"
"06"	"Goma"
"07"	"Gwana"
"08"	"Hidim"
"09"	"Istan"
"10"	"Kabi"

Steps for recoding >12 responses, continued

6. When all of the values have been entered, select **OK** to exit the Recode dialog box.

   Sit_num values from "11" – "19" will be coded as (.) or missing in the *SiteName* field.

7. Select the **Recode** command from the Analysis menu tree to recode the rest of the sites.

8. Select *sit_num* as the **From** variable and *tempSite1* as the **To** variable.

9. Using the following information, recode *tempSite1:*

Sit_num	tempSite1
"11"	"Karanda"
"12"	"Loma"
"13"	"Maka"
"14"	"Mindi"
"15"	"Mura"
"16"	"Mustubini"
"17"	"Nabo"
"18"	"Nkula"
"19"	"Tapanda"

   Sit_num values from "01" – "10" will be coded as (.) or missing in the *tempSite1* field.

10. Using the **IF** command, set the value of *SiteName* equal to the value of *tempSite1* where *SiteName* is missing (.) Your code should appear as below.

```
IF SiteName=(.) then
ASSIGN SiteName=tempSite1
END
```

11. Save your program again.

12. **Write (Export),** selecting the **Replace** output method to the *Allclean* table.

13. **Read (Import**) the *Allclean* table.

14. Verify recode of *SiteName* using **Frequency**.

Steps for recoding >12 responses, continued

Note that in Epi Info, you cannot perform a frequency after completing complex recodes without first writing out the file and then reading it again. In the example above, if you perform a frequency of *SiteName* to check your work, you will receive an error notification requiring you to exit Epi Info. Therefore, it is always best to complete the recodes (taking care to not overwrite the original data in any case!) and then write and re-read the file to check your work.

Try it yourself!

Activity 4, Create a Text Variable

In order to perform trend analysis by year, we need to have the year of the visit date in a separate field. First, make sure that no records are missing a *vst_date* value. Next, **Define** the variable Year and then **Assign** *Year* to a four-digit text value for the client visit date.

Creating a Data Analysis File

Saving data sets for analysis purposes

Once all recodes have been completed, you need to save a data set for analysis purposes. Data sets for analysis purposes should include only the variables that you will use in your analysis.

1. Click **Write (Export)** in the tree command box.

   Verify that the output format is Epi 2000.

2. Type *C:\ANC_Suri\Analysis\ANCAll.mdb* into the **File Name** prompt.

3. Type *Analysis* as the table name into which data will be saved with the new variables.

Saving data sets for analysis purposes, continued

4. Select the following fields from the **Variables** box:

   - Pt_key
   - Occup1
   - Par
   - Grav
   - Region
   - AgeGroup
   - District1
   - SiteName
   - Education1
   - Residence1
   - MarStatus1
   - HIV
   - TPHA
   - RPR
   - Year

5. Make sure **Replace** is selected.

6. Click **OK**.

Don't forget to save the program code you created in the program editor file, as well, in case there were errors in the recoding process. Saving the program will ensure that you don't have to type the recode information again.

Checking your work

Once completed, read in the new *ANCAll* dataset and the *Analysis* table. Perform several frequencies of the variables to ensure that all of the recodes were successful.

Exercise 9
Performing Descriptive Analysis

Overview

What this exercise is about

According to the data analysis plan, characteristics of the sample population and their HIV and RPR prevalence for 2002 (including 95% confidence intervals) are important survey outcomes. To calculate these simple descriptive statistics, we will use Epi Info Analysis. These data can then be used to produce charts and graphs. Results from this section of the analysis will be useful when developing a national report describing the HIV sentinel survey results.

What you will learn

At the end of the exercise, you will be able to:

- Perform simple descriptive analyses:
  - calculating means, medians and frequencies describing sample group characteristics
  - generating frequencies of HIV prevalence by sub-groups
  - using the tables command to generate sub-group analyses.
- Produce graphs, charts and maps that illustrate key findings.
- Modify the properties of graphs, charts and maps.

Starting location

Analysis, C:\ANC_Suri\Analysis\ANCall.mdb:Analysis

Generating Sample Population Statistics

Variables of interest

At the end of unit 1 (Exercises 1–8), we developed a data analysis plan, beginning with the calculation of simple descriptive statistics for the sample population, including the HIV and RPR prevalence with associated 95% confidence intervals. Variables of interest are:

- site
- district
- age group
- marital status
- educational level
- residence
- gravida
- parity
- occupation.

Generating frequencies, reporting minimum (min) and maximum (max) values, and calculating means and medians are the primary methods for generating these statistics.

Frequencies in the Sample Population

Generating frequencies

Begin by generating a frequency of the number of pregnancies per woman, including the frequency of women for whom this is the first pregnancy.

1. Click **Read (Import)** from the command tree on the left side of the Analysis window.

2. Change the project to the *C:\ANC_Suri\Analysis\ANCall.mdb* project file.

3. Select **Show All** to see and select the *Analysis* Table.

4. Click **OK**.

5. Select only those records where *Year = "2002"*. You should have 6 604 records in your 2002 database sub-set.

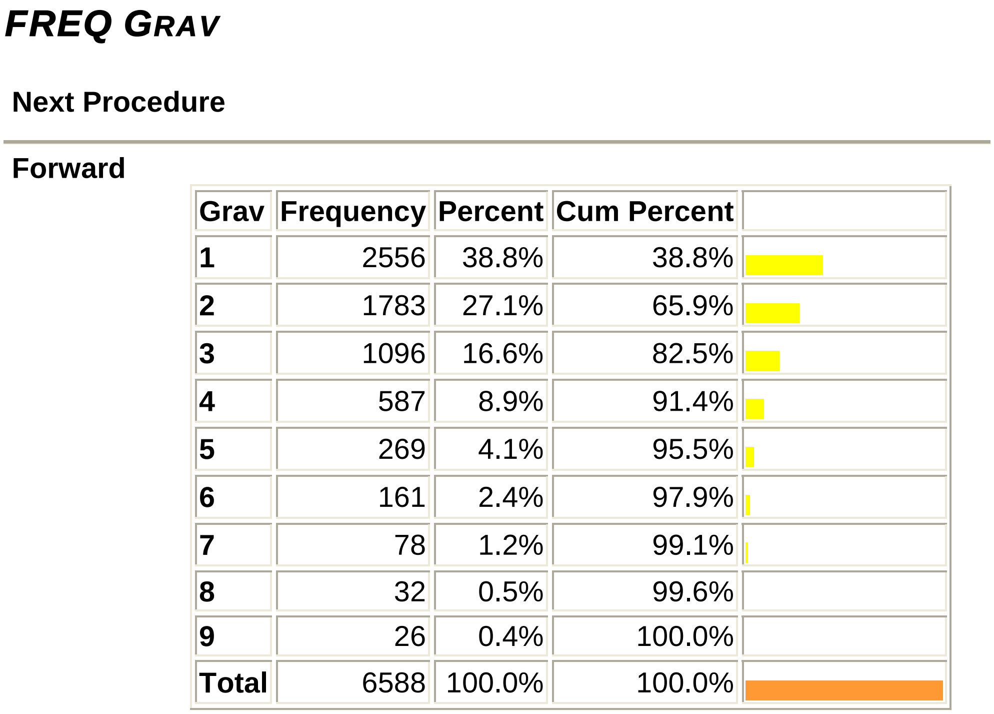
Generating frequencies, continued

6. Select **Frequencies** from the command tree.

7. Select *Grav*.

8. Click **Settings**. In the Settings box, ensure that the **Include Missing** option is NOT selected, and then click **OK**.

9. Click **OK** in the **FREQ** box. A frequency of the number of pregnancies per woman appears.



FREQ GRAV

Next Procedure

Forward

Grav	Frequency	Percent	Cum Percent	
1	2556	38.8%	38.8%	
2	1783	27.1%	65.9%	
3	1096	16.6%	82.5%	
4	587	8.9%	91.4%	
5	269	4.1%	95.5%	
6	161	2.4%	97.9%	
7	78	1.2%	99.1%	
8	32	0.5%	99.6%	
9	26	0.4%	100.0%	
Total	6588	100.0%	100.0%	

Notice that our total number of women (6 588) is less than the total number of records selected (6 604) because we have removed from our analysis women with missing values for the *Grav* variable.

As you may also have noticed, the process for doing frequencies during data analysis is the same as during data cleaning. We will provide examples of how to present and interpret these frequencies below.

Min, Max, Median and Mean Values in the Sample Population

It is generally useful to describe the min, max, median and/or mean values of characteristics in your population that are measured on a continuous scale, like number of pregnancies per woman.

- **Min value** is that observation that is the lowest in the dataset for a particular variable.

- **Max value** is that observation that is the highest in the dataset for a particular variable.

- **Median value** is the observation that indicates the point where half of the observations are less than, or greater than, the value. Median values are unaffected by extreme high or low values. The median is often also called the "50th percentile."

- **Mean value** is the sum of all the values added together divided by the total number of values. The mean can be affected by extreme values. Therefore, it is important to consider the effect of individual values when reporting the mean. The mean is often also called the "average."

Obtaining minimum, maximum, median and mean values

To obtain minimum, maximum, median and mean values describing the number of pregnancies among the women enrolled in the survey in 2002, we can follow the example below.

1. Select the **Means** command from the command tree.

2. Select *Grav* from the **Means of** drop-down menu.

3. Click on the **Settings** command button to deselect the graphics, percent and output tables, since we are interested in the overall means for the Count variable and not the individual percents and graphics for the *Grav* variable.

4. Ensure that the **Include Missing** option is <u>NOT</u> selected. Once this is done, click **OK** on **Settings**.

5. Click **OK** on the **Frequency** dialog box.

Obtaining minimum, maximum, median and mean values, continued

The following output should provide you with the information you need to summarise the number of pregnancies per woman among the women participating in the 2002 survey:

Obs	Total	Mean	Variance	Std Dev
6588	15105.0000	2.2928	2.2481	1.4994

Minimum	25%	Median	75%	Maximum	Mode
1.0000	1.0000	2.0000	3.0000	9.0000	1.0000

Again, notice that the total number of women described in this analysis is 6 588. This is the same group of women described in the frequency analysis above.

Summarising the Amount of Missing Data

Reporting the number and percent of individuals with missing data

As a general rule, all descriptive statistics should be performed and interpreted on the group of individuals with non-missing values for the characteristic of interest, which is why we recoded the missing values in Exercise 8 and why we made sure that we excluded them from these two previous analyses.

It is important, however, to report the overall number and percent of individuals with missing information, because this allows your audience to gauge how reliable or generalisable the data are to the population under study.

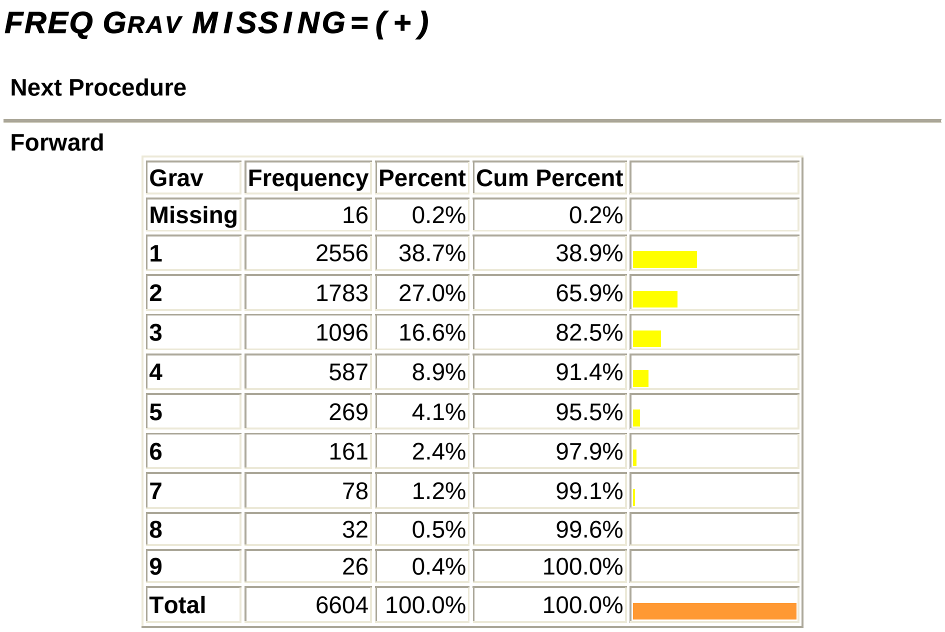
Try it yourself!

Activity 1, Calculate Number and Percent

To calculate the number and percent of women participating in the 2002 survey who were missing information on gravidity, rerun the frequency of *Grav*. This time, however, we want to select the **Include Missing** option in the **Settings** box.

Activity 1, Calculate Number and Percent, continued

The frequencies of number of pregnancies per woman, with the missing values included, appear as follows:



FREQ GRAV MISSING=(+)

Next Procedure

Forward

Grav	Frequency	Percent	Cum Percent	
Missing	16	0.2%	0.2%	
1	2556	38.7%	38.9%	
2	1783	27.0%	65.9%	
3	1096	16.6%	82.5%	
4	587	8.9%	91.4%	
5	269	4.1%	95.5%	
6	161	2.4%	97.9%	
7	78	1.2%	99.1%	
8	32	0.5%	99.6%	
9	26	0.4%	100.0%	
Total	6604	100.0%	100.0%	

This time, our total (6 604) equals the records selected because we have included our missing values. Notice that 16 women, or 0.2% of our sample, were missing information on gravidity.

Presenting and Interpreting Frequencies, Min, Max, Median and Mean Values

As mentioned above, the statistics that we report and interpret for a certain characteristic should be based on the group of individuals with non-missing values for that characteristic. If the data are not missing because of some systematic reason that would introduce bias into our survey, then analysing the non-missing data is the best method of analysing and presenting the survey data. We will also report the overall number and percent of individuals with missing information.

Presenting frequencies

An example of a table that presents frequencies among the non-missing cases and also reports the amount of missing data for quality purposes is shown below. Note that **this is not a table created directly from Epi Info**, but rather an example of a table suitable for presenting in a report, in which we have summarised the results of the analyses with and without the missing values. Which values in the table and in the interpretation are drawn from the analyses among the non-missing cases, and which values from the analysis that included the cases missing gravidity?

Table 1. Number of total pregnancies among women participating in the ANC survey, Suri 2002.

Number of lifetime pregnancies	Number of women	Percent
1	2 556	38.8%
2	1 783	27.1%
3	1 096	16.6%
4	587	8.9%
5	269	4.1%
6	161	2.4%
7	78	1.2%
8	32	0.5%
9	26	0.4%
Missing/Unknown	16	
Total	**6 604**	**100.0%**

Interpreting the data presented in Table 1

Interpretations of the frequency, min, max, median and mean data for gravidity are as follows:

- In 2002, 6 604 ANC clients were screened as part of the ANC HIV sentinel surveillance survey.
- For nearly 39% (2556/6588) of women, this pregnancy was the first.
- The number of pregnancies ranged from 1 to 9 lifetime pregnancies, with half of the women having had 2 or fewer pregnancies. The average number of pregnancies among survey participants was 2.3.
- All descriptions of gravidity are reported based on the women with valid information on pregnancy history. This information was missing for 16 (0.2%) of the survey participants.

Try it yourself!

Activity 2, Generate Summary Statistics

Generate summary statistics (frequencies or min/max/median/mean where appropriate) for District, Age Group, Marital Status, Educational Status, Residence, Parity and Occupation for the 2002 data.

Note that when reporting frequencies it is often helpful to the reader to put the percent value with the numerator and denominator in parentheses. For example, when we reported on gravidity, the number of women on their first pregnancy was 38.8% (2556/6588) of survey participants with valid information for this characteristic.

Write a sentence or sentences that describe your results for the following categories. For each variable, remember to calculate frequencies among the cases with non-missing values and also to report the number and percent of cases with missing values.

- **District** – Describe the percentage of women in each district sampled.

- **Age Group** – Describe the percentage of women in your sample who were between 15 and 24 years of age at the time of the survey.

- **Marital Status** – Identify the percentage of women in the largest category of marital status.

- **Educational Status** – List the percentage of women who had completed primary school at minimum. Qualitatively compare it with the percentage of women who completed no schooling.

Activity 2, Generate Summary Statistics, continued

- **Residence** – Describe the percentage of women included in the sample who live in urban areas vs. rural areas.

- **Parity** – Describe the number of live births per woman.

- **Occupation** – Identify the two most common occupations and the percent of women who list those as their occupations.

Describing Sample Size Per Survey Site

Describing the survey site

In all of the previous analyses, we described characteristics of our survey participants. We performed these analyses by applying the Epi Info analysis functions directly to our *Analysis* table. You may not have realised that we were able to do this because in our *Analysis* table, each line of data in our database represents one survey participant, which coincides with the person we were describing.

In our data analysis plan, however, we noted that we also wanted to describe the survey sites. It is good practice to report the sample size achieved for each survey site, as well as the mean and range of number of participants enrolled in each of the survey sites.

"Sample size per survey site" is a characteristic of a survey site, however, and not of an individual woman. This means that in order to perform this analysis we will first need to create an intermediate data table containing the survey sites and their sample sizes, where each record represents a survey site. This is easy to do in Epi Info and is illustrated in the example below.

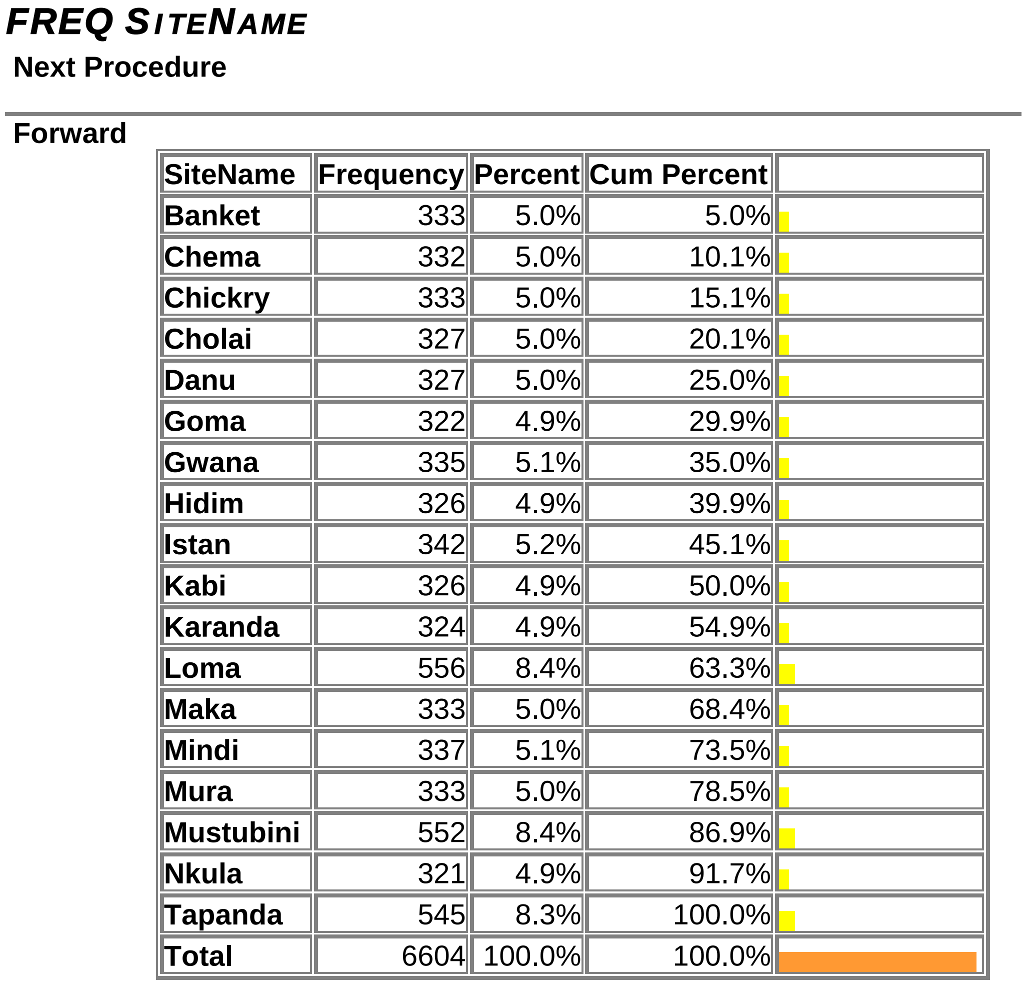
Generating a frequency

We will begin by generating a frequency of the number of women sampled per site in 2002.

1. Make sure that the Analysis table is open and that only the 2002 records are selected. You should have 6 604 records in your 2002 database sub-set.

2. Select **Frequencies** from the command tree.

3. Select *SiteName*. A frequency of the number of women sampled per site appears.



FREQ SITENAME

Next Procedure

Forward

SiteName	Frequency	Percent	Cum Percent	
Banket	333	5.0%	5.0%	
Chema	332	5.0%	10.1%	
Chickry	333	5.0%	15.1%	
Cholai	327	5.0%	20.1%	
Danu	327	5.0%	25.0%	
Goma	322	4.9%	29.9%	
Gwana	335	5.1%	35.0%	
Hidim	326	4.9%	39.9%	
Istan	342	5.2%	45.1%	
Kabi	326	4.9%	50.0%	
Karanda	324	4.9%	54.9%	
Loma	556	8.4%	63.3%	
Maka	333	5.0%	68.4%	
Mindi	337	5.1%	73.5%	
Mura	333	5.0%	78.5%	
Mustubini	552	8.4%	86.9%	
Nkula	321	4.9%	91.7%	
Tapanda	545	8.3%	100.0%	
Total	6604	100.0%	100.0%	

Generating a frequency, continued

Note that the Nabo clinic, site "17," has been excluded from analysis. As you may recall from the first part of the course, Nabo had laboratory testing problems.

Categorising sites based on sample size

In this example, to obtain minimum, maximum and mean values describing the number of woman sampled per site in our data, we could calculate these descriptors for all 18 sites. Because there are three sites with large sample sizes (to oversample women <25 years of age), however, we might also want to consider generating statistics for the 15 small sites separate from the three large sites. If we averaged all of the 18 sites together, we might overestimate the mean sample size at most sites.

To obtain minimum, maximum and mean sample sizes for the 15 smaller sites, we can follow the example below.

1. Choose the **Select** command from the command tree and select those sites where the sample size is more closely related. You can select all 15 sites individually by listing their names, or, more simply, select all sites other than the three large sites. Code for the select statement appears below:

   SELECT SiteName <>"Loma" AND SiteName<>"Mustubini" AND SiteName<>"Tapanda"

2. Select the **Frequencies** command from the command tree. There should be 4 951 records in your data set.

3. Select *SiteName* from the variable list.

4. Type the table name *SiteNameCount* into the **Output to Table** prompt.

   This table will store the *Count* variable or the number of women sampled by site (*SiteName*).

5. Click **OK**.

Creating intermediate data tables

The introduction mentioned that we would need to create an intermediate data table containing the survey sites and their sample sizes as the first step of describing the characteristics of the survey sites analysis. We have just created this intermediate table. Next, we will need to read it back into Epi Info to make it the active table, or, in other words, the table that we are analysing.

6. **Read (Import)** the *SiteNameCount* table from the **Project** *C:\ANC_Suri\Analysis\ANCall.mdb.*

7. Select the **Means** command from the command tree.

8. Select the **Means of** the *COUNT* variable.

   You can click on the **Settings** command button to deselect the graphics, percent and output tables since we are interested in the overall means for the Count variable and not the individual percents and graphics for each site. Once this is done, click **OK**.

9. Click **OK**.

Summarising sample sizes across sites

The following table should provide you with the information you need to summarise the 15 sites and their sample sizes:

Obs	Total	Mean	Variance	Std Dev
15	4951.0000	330.0667	34.9238	5.9096

Minimum	25%	Median	75%	Maximum	Mode
321.0000	326.0000	332.0000	333.0000	342.0000	333.0000

Interpretation of the frequency, min, max and mean data are as follows:

- In 2002, 6 604 ANC clients were screened as part of the ANC HIV sentinel surveillance survey.
- Fifteen of the 18 sites collected between 321 and 342 samples, with an average of 330 women sampled per site.

Try it yourself!

Activity 3, Describe the Sample Sizes for the Three Large Sites

Calculate the min, max and mean data for the three large sites. Write a statement below that describes your findings.

Hint: Remember to re-**Read** the Analysis table and **Select** year= “2002” again!

Understanding Confidence Intervals

Confidence intervals

A *confidence interval (CI)* gives an estimated range of values, which is likely to include an unknown population parameter, the estimated range being calculated from a given set of sample data. If independent samples are taken repeatedly from the same population, and a confidence interval calculated for each sample, then a certain percentage (confidence level) of the intervals will include the unknown population parameter. Confidence intervals are usually calculated so that this percentage is 95%, but we can produce 90%, 99%, and 99.9% confidence intervals for the unknown parameter.

A 95% confidence interval means that if the study were repeated 100 times, 95 out of 100 times the CI would contain the true measure of disease.

The width of the confidence interval gives us some idea about how uncertain we are about the unknown parameter. A very wide interval may indicate that more data should be collected before anything very definite can be said about the parameter.

Confidence intervals are more informative than the simple results of hypothesis tests (where we decide 'reject H0' or 'don't reject H0'), since they provide a range of plausible values for the unknown parameter.

Confidence limits

Confidence limits are the lower and upper boundaries/values of a confidence interval, that is, the values that define the range of a confidence interval. The upper and lower bounds of a 95% confidence interval are the 95% confidence limits. These limits may be taken for other confidence levels–for example, 90%, 99% and 99.9%.

Using a hypothetical example, it may be reported that “The estimated number of people living with HIV (prevalence) among ANC attendees was 18.6%, with a 95% CI of 12.9-24.0.” This means that the study investigators are 95% sure that the true prevalence lies somewhere between the two confidence limits of 12.9% and 24.0%.

If there were 1 000 ANC attendees under study, it would be reported that approximately 186 (18.6%) of them were living with HIV. It is most accurate to say that the study investigators are 95% sure that between 129 (12.9%) and 240 (24.0%) of the ANC attendees in the study were living with HIV.

Calculating Prevalence Confidence Intervals

Calculating HIV prevalence

HIV sero-prevalence (*P*) and the associated 95% CIs are the primary outcomes of interest when analysing ANC survey data. HIV prevalence is calculated as:

$$P = x/n$$

where *x* is the total number of persons testing positive for HIV and *n* is the total number of specimens tested at a given site or among sub-group members (e.g., 20-24-year old ANC patients).

Multiplying the proportion, *P*, by 100% will express HIV prevalence as the percentage positive. For example, if 93 of 500 specimens at a sentinel site are HIV-positive, the HIV prevalence at that ANC site is 18.6% (93/500 x 100%).

Calculating HIV prevalence, continued

To calculate the Exact Binomial CI for sero-prevalence estimates in Epi Info, the Unadjusted CI formula is used:

$$\{P \pm \{1.96 \sqrt{[(1-P)\ P]/n)}\}\} * 100$$

where P = prevalence
n = total number of specimens tested

Example: In the above example,
$P = (93/500) = .186$
$n = 500$

thus: $95\%\ CI = \{0.186 \pm \{1.96 \sqrt{[(1-.186)\ .186]/500)}\}\} \times 100$
$= (0.186 \pm .034) \times 100$
$= 15.2\%$ to 22.0%

The lower bound (15.2%) and upper bound (22.0%), or confidence limits, in the unadjusted CI are similar to the Exact Binomial CI when sample sizes are large.

Calculating HIV prevalence and CI

To calculate HIV prevalence and the CI, we use the Epi Info frequency command:

1. Click **New** on the Program Editor menu bar in Analysis.

   This clears the other commands that have been saved and/or executed.

2. Locate the *C:\ANC_Suri\Analysis\ANCall.mdb* project file or type it into the project prompt box. Select *Analysis* as the Table Name.

3. **Select** Year = “2002”.

4. Select the **Frequencies** command in the command tree to develop a 2 by n table.

Calculating HIV prevalence and CI, continued

As noted previously, HIV prevalence can also be calculated using the **Tables** command. Currently, however, **Tables** does not provide a confidence interval estimate in Epi Info. For frequencies that involve more than 2 by n cross-tabulations, you must use **Tables** and calculate the CIs by hand or use a different software tool if appropriate.

5. Select *HIV* as the **Frequency of** variable.

6. Select *SiteName* as the **Stratify by** variable.

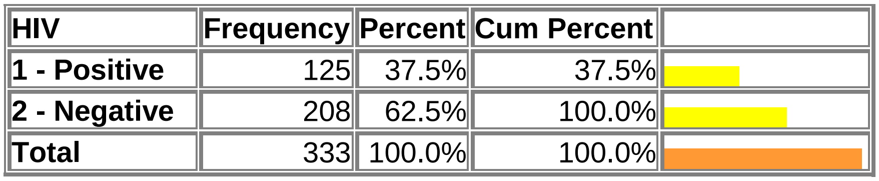
7. Click **OK**.

Example of HIV frequency table

An HIV frequency for each site should be produced, as shown below in the Banket example.

HIV, SiteName=Banket

Forward



HIV	Frequency	Percent	Cum Percent	
1 - Positive	125	37.5%	37.5%	▬
2 - Negative	208	62.5%	100.0%	▬
Total	333	100.0%	100.0%	▬

Example of exact Binomial CI table

Below the Epi Info table of HIV prevalence by SiteName is the Exact Binomial 95% CI:

95% Conf Limits		
1 – Positive	32.4%	43.0%
2 – Negative	57.0%	67.7%

To summarise HIV prevalence in Banket, we estimate HIV prevalence to be 37.5% among ANC clients aged 12-49. Further, we have 95% confidence that the true HIV prevalence among the eligible population in the catchment area is between 32.4% and 43.0%.

Try it yourself!

Activity 4, Calculate Overall HIV Prevalence and 95% Confidence Interval

Calculate the overall HIV prevalence and 95% confidence interval among ANC site attendees sampled in the 2002 survey. Summarise these results.

In addition, summarise the results of HIV prevalence by site, highlighting the sites with the highest and lowest HIV prevalence and their 95% CIs.

Interpreting Differences Using Confidence Intervals

In Activity 3, we calculated HIV prevalence and a 95% CI for each of the 18 sites. Remember that the HIV prevalence calculated for each site is based on a sample of women, which is a small proportion of the population of child-bearing women who access prenatal care at the clinic or who live in the area served by the clinic. The HIV prevalence that we calculated is our best estimate of the true HIV prevalence in the entire population of child-bearing women attending the ANC clinic. Because we haven't tested all women in the population, however, we must reflect some uncertainty in our estimate. This uncertainty is reflected in the confidence interval.

We applied this concept of uncertainty when we interpreted the HIV prevalence for the Banket site. We noted that, although we estimated the HIV prevalence to be 37.5%, "we have 95% confidence that the true HIV prevalence among the eligible population in the catchment area is between 32.4% and 43.0%."

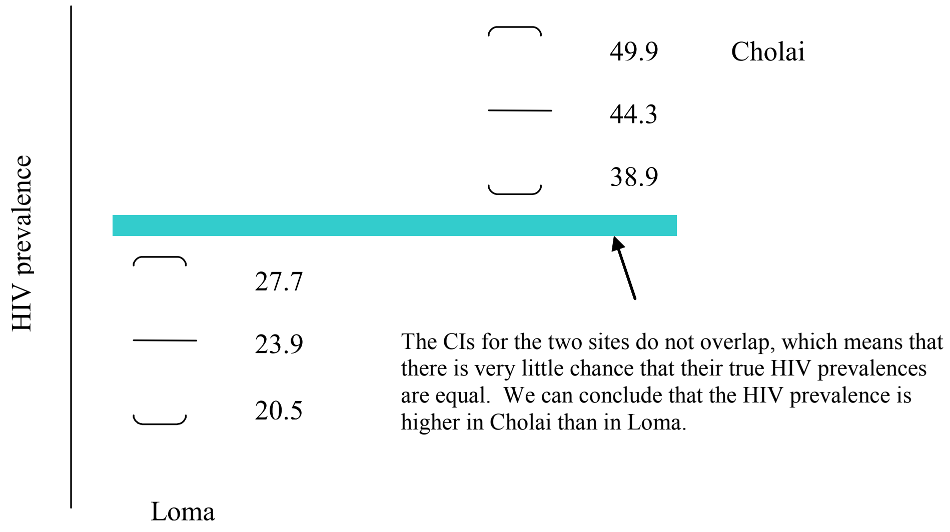
This concept of uncertainty is very important when trying to determine how different the HIV prevalence is among different groups. In Activity 3, we saw that Cholai and Loma had the highest and the lowest HIV prevalence, with corresponding 95% confidence intervals of (38.9%, 49.9%) and (20.5%, 27.7%), respectively.

Interpreting Differences Using Confidence Intervals, continued

The 95% CI for Cholai tells us that there is a chance that the HIV prevalence could be as low as 38.9% in the population served by the site. The 95% CI for Loma tells us that there is a chance that the HIV prevalence could be as high as 27.7% in the population. So, is the HIV prevalence really higher in Cholai than in Loma?

Using pictures or graphics to help interpret differences

The clearest, easiest way to decide this is to draw a picture. If you have access to Excel or to graphing software, you can use these tools. However, drawing a quick picture by hand is generally the best way to answer the question.



The CIs for the two sites do not overlap, which means that there is very little chance that their true HIV prevalences are equal. We can conclude that the HIV prevalence is higher in Cholai than in Loma.

Try it yourself!

Activity 5, Compare the HIV Prevalence of Banket and Chema

From Activity 3, we saw that the HIV prevalence in Banket's sample of ANC attendees was lower than in the sample from Chema (37.5% versus 38.9%). Draw a picture that includes the 95% CIs for these two sites, to determine if the HIV prevalence for the populations served by these appear to differ.

Activity 6, Calculate HIV Prevalence for 2002

Follow the steps in the previous pages to calculate HIV prevalence in 2002 for the following:

- **District** – Which district has the highest overall HIV prevalence? The lowest?
- **Marital Status** – Which group (single, widowed, etc.) experienced the highest HIV prevalence?
- **Educational Status** – Does educational status appear to affect HIV prevalence?
- **Residence** – Are rural women more likely than urban women to be HIV positive?
- **Gravidity and Parity** – Describe HIV prevalence by gravidity and parity. Are women experiencing their first pregnancy more or less likely to be HIV positive?
- **Occupation** – Do any of the occupation categories seem to experience higher HIV prevalence than other categories?
- **Age Group** – Describe the overall HIV prevalence trend by age group. Which age group has the highest HIV prevalence? The lowest?

  Be sure to save your program code before moving on to the next section.

Graphing Output

Charts and graphs as tools for communicating data

Epidemiologists use HIV prevalence tables to display patterns in the data, but charts and graphs are often better tools for communicating information at a glance.

Charts and graphs as tools for communicating data, continued

When creating charts and graphs, you always compare your graph to a table of values on which the graph is based. Make sure that you have correctly specified the attributes of your graph in the Epi Info interface.

When you present your graph, always make sure to label it with a title that describes the contents of the graph including the x and y axis fields and the type (total count, percent, etc.).

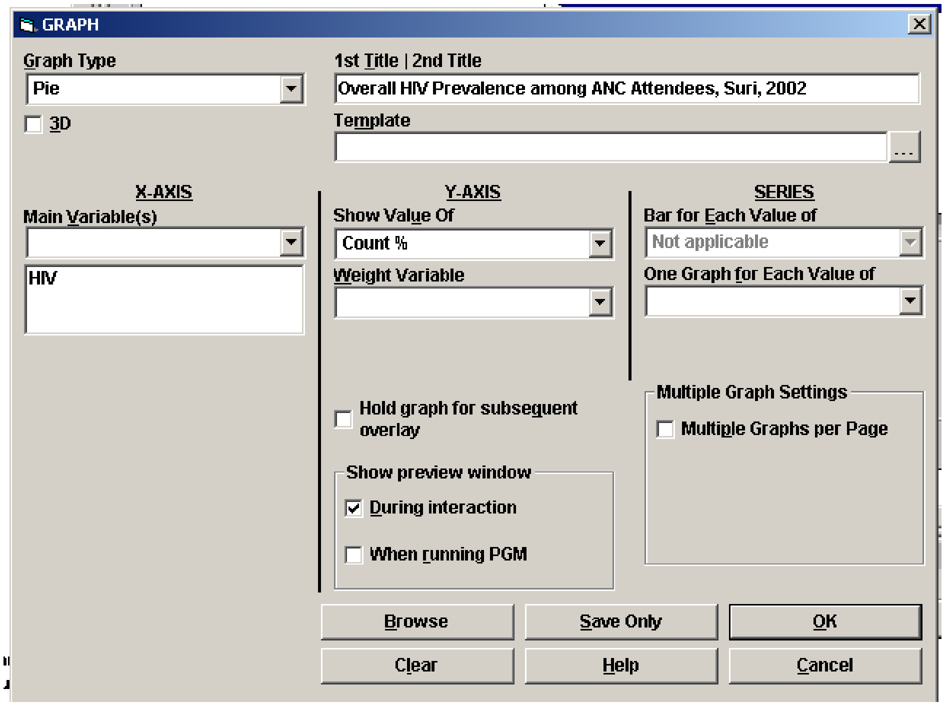
Creating Pie Charts

Steps to generate pie charts

To generate a pie chart of overall HIV prevalence for the sites,

1. Click on the **Graph** command. Select or type the following information into the command box:

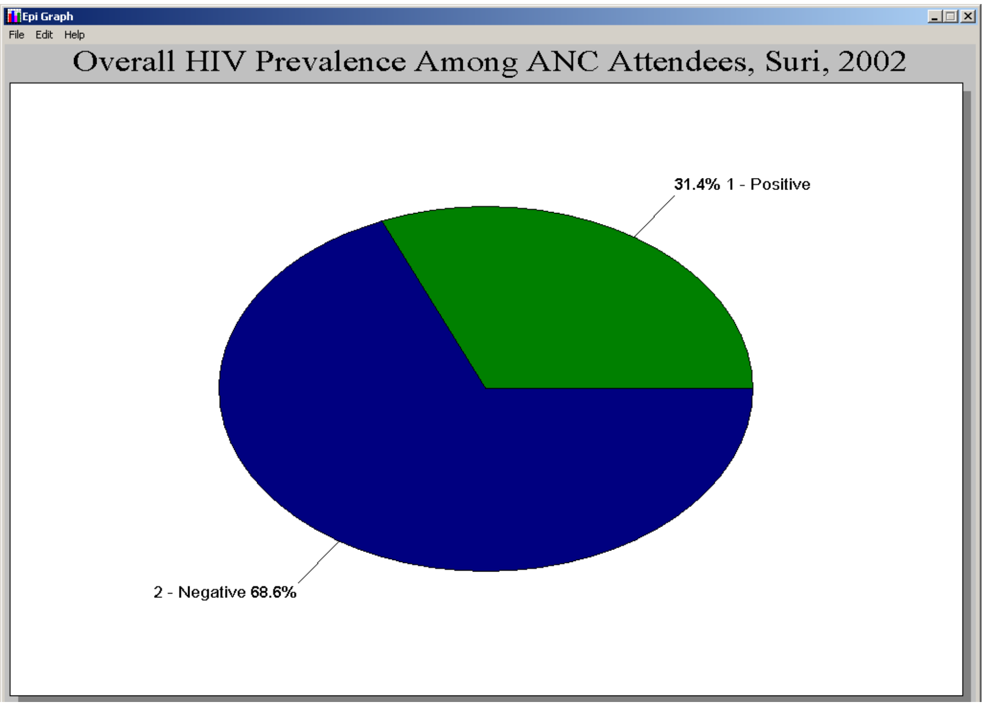
   Graph type: Pie
   X-Axis Main Variable: HIV
   Y-Axis Show Value Of: Count %



Graph Type: Pie
□ 3D
1st Title | 2nd Title: Overall HIV Prevalence among ANC Attendees, Suri, 2002
Template:
X-AXIS Main Variable(s): HIV
Y-AXIS Show Value Of: Count %
Weight Variable:
SERIES Bar for Each Value of: Not applicable
One Graph for Each Value of:
□ Hold graph for subsequent overlay
Show preview window: ☑ During interaction □ When running PGM
Multiple Graph Settings: □ Multiple Graphs per Page

Steps to generate pie charts, continued

2. Click **OK**. A pie graph showing HIV prevalence will appear in the EpiGraph window.



3. Right-click on the graph to bring up a menu where you can manipulate the appearance of the graph or double-click with the left mouse button to pull up a customisation dialog box.

   Use the customisation dialog box as follows:

   - Under the General tab, add:
     - Main Title: Overall HIV Prevalence Among ANC Attendees
     - Sub-title: Suri, 2002
   - Under the fonts tab, select Arial as the default font for all values
   - Under the Style tab:
     - Change HIV Positive to be red
     - Change HIV Negative to be yellow.

4. Click **OK**.

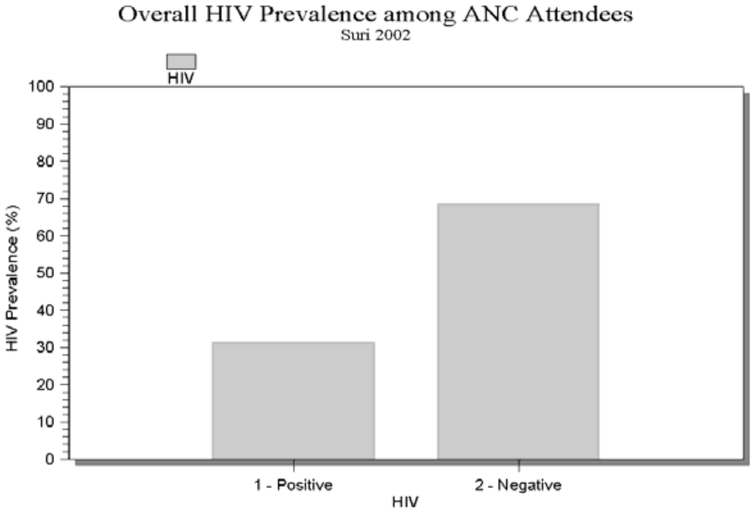
5. From the **File** menu, select **Save & Exit** from the EpiGraph once the changes have been made.

   Your graph should be displayed in the output window of Analysis.

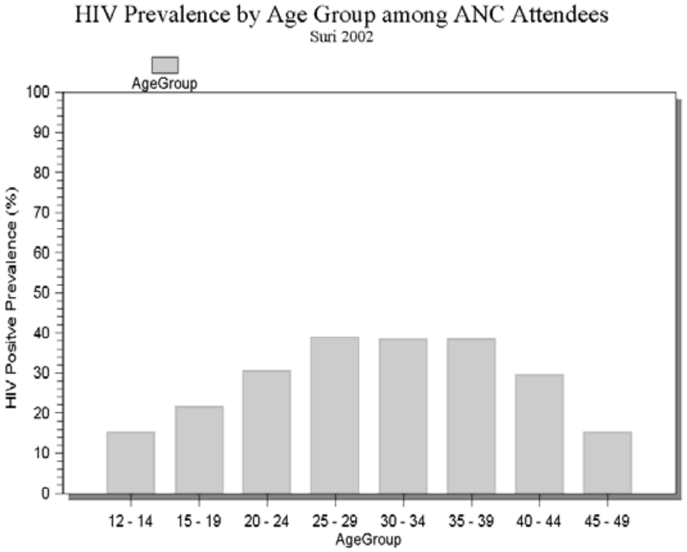
Creating Bar Charts

Bar charts

As with the pie graph, a single proportional outcome (i.e., overall HIV population prevalence) can be graphed as a bar chart using the **count %** on the Y axis.



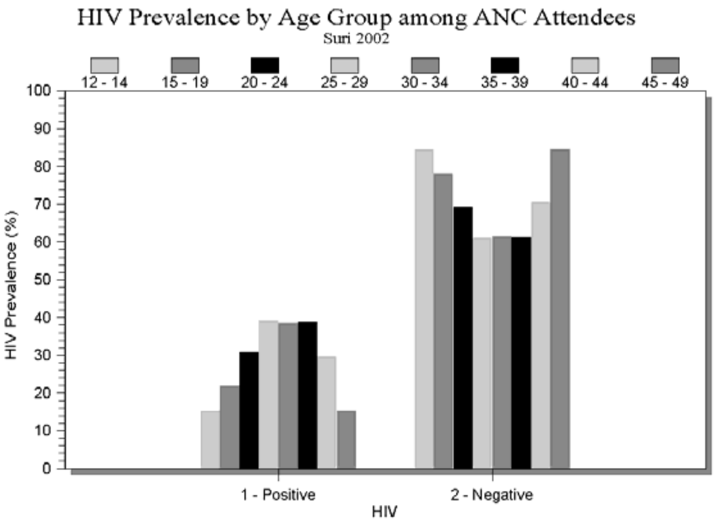
If you are graphing proportional data with multiple outcomes, such as HIV positive prevalence in each age group, we must first create a new variable that we will use as a "weight" in our graphs. After creating our new variable, the calculation of HIV prevalence can be graphed for each age group as shown below. The count % is no longer necessary because we will be graphing the average of our weight variable instead.



Bar charts, continued

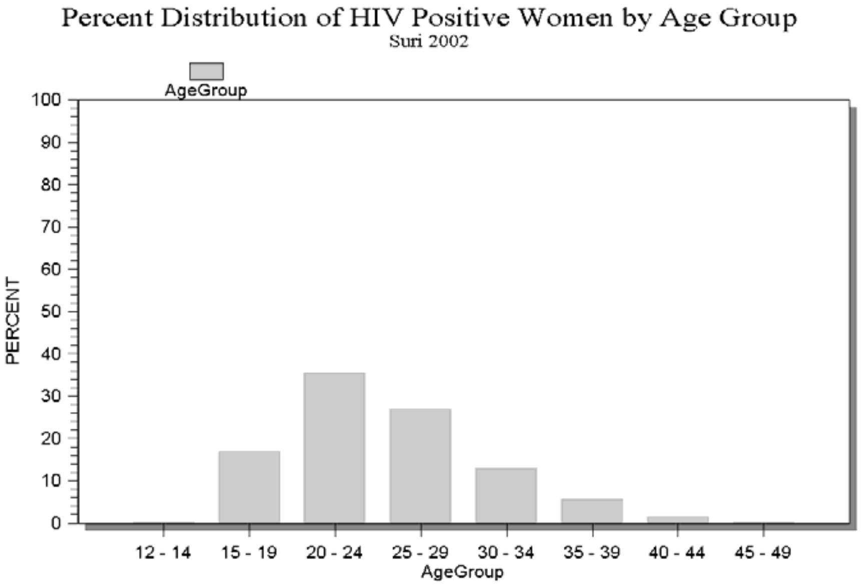
Note that if a new HIV prevalence variable is not created, the bar graph would either show:

1. HIV-positive and -negative prevalence on the same graph. This is cumbersome and unnecessary.



Or:

2. The total number of women infected with HIV for that age group out of all women tested (i.e., the column percent) rather than the total number of women HIV-infected out of the total number of women in the age group (i.e., the row percent). This example is shown below.



Does your graph show what you intended?

If you are confused about what your graph is showing, look back at the frequencies and tables to figure it out. Make sure the graph is showing what you intend to show. In the Epi Info **Customization Dialogue Box**, you can also show the data table for the graph by clicking on that option in the **General** Tab.

Steps to create bar charts

To graph HIV-positive prevalence in each age group in a bar chart, follow the steps below:

Step 1 – Create a variable whose mean value will yield HIV prevalence.

1. **Cancel Select** to return to the full dataset with 18 596 records.

2. **Define** the variable *HIVPos.*

3. Create an If/Then statement, using the **If** command, to assign HIV-positive values a value of 100 and to assign HIV-negative values a value of 0. Your IF/THEN statement should appear in the program editor as shown below:

```
IF HIV="1 – Positive" THEN
    ASSIGN HIVPos=100
ELSE
    ASSIGN HIVPos=0
END
```

4. Check your recode by running a frequency on your new variable *HIVPos*. If your results are not correct, please repeat instruction 2.

5. Once you have verified your changes were made correctly, **Write (Export)** and **Replace** the *Analysis* table. This will ensure that *HIVPos* is a permanent field in the *Analysis* table.

6. **Read** the *Analysis* table.

7. **Select** the year "2002" to continue analysis on the 2002 ANC dataset.

Steps to create bar charts, continued

Step 2 – Create a bar graph of HIV prevalence in each age group.

To create a bar graph of HIV-positive prevalence in each age group,

1. Click on the **Graph** command. Select or type the following information into the command box:

   Graph type: Bar
   X-Axis Main Variable: AgeGroup
   Y-Axis Show Value Of: Average
   Weight: HIVPos

2. Click **OK**.

3. Double-click on the Y-axis label to change the label from HIVPos to HIV Prevalence (%).

4. Right-click on the graph, and select **Customization Dialogue**….

   - Under the General tab:
     - Main title: Percentage HIV-Positive by Age Group Among ANC Attendees
     - Sub-title: Suri, 2002
   - Under the Axis tab, change the Y-axis min/max values to 0 and 100.
   - Under the Font tab, change the default font to Arial.
   - Under the Style tab, change bar colors from green to blue.

5. Click **OK**. The changes will be applied.

Steps to create bar charts, continued

6. Click on **Save & Exit** from the EpiGraph Menu once the changes have been made. Your graph should be displayed in the output window of **Analysis**.

Analysis results in the Display Window are in an HTML file format and cannot be changed in Epi Info. Pictures of the graphs are created in .jpg file format and also cannot be modified. However, both outputs in the display window can be copied directly to other applications, such as PowerPoint and Microsoft Word, by using the mouse and standard tools for copying, (Ctrl+C) and pasting (Ctrl+V).

Copying and pasting results from Epi Info to another application is useful when creating reports.

The results of each Analysis session displayed in the Display window are saved as an HTML file in the Project's File Folder as OutXXX.htm. You can open these results in any HTML browser to review previous analyses.

7. Save your program code with a suitable name.

Try it yourself!

Activity 7, Create a Bar Graph

Using the set of instructions from Step 2, create a bar graph of HIV-positive prevalence by residence.

Use Maps to Visualize Your Data

Using Epi Map

Mapping prevalence and other types of data is another way to summarise your results. You can use **Epi Map** to do this. Epi Map takes data collected using Epi Info and displays the values on a map.

To correctly draw the necessary geographical features, Epi Map needs boundary files that contain information about the geographical boundaries, labels (names) and other cartographic (map) data. Epi Map uses a boundary file format called Shape files, which was developed by a company called Environmental Sciences Research Institute (ESRI). The shape file format is used more than any other boundary file format and has become the standard in the industry.

Using Epi Map, continued

For this exercise, we will map the data from the previous exercises. We will explore two different ways to map the data: first using Analysis, and second using Epi Map. Both methods will result in the same map.

Preparing Data for Mapping

Our objective is to map the HIV prevalence data of each district. In order to do this, a table containing the district name and the prevalence data has to be created.

Try it yourself!

Activity 8, Construct a Data Table for Epi Map

Within the *ANCAll* project, we will create a table called *HIV_prev* containing the figures for HIV prevalence by District Name using the **Summarize** command.

To create a variable representing HIV prevalence:

1. Click on **Summarize** in the command tree.

2. Select Average from the **Aggregate** drop-down box.

3. Select *HIVPos* from the **Variable** drop-down box.

4. Type *Prevalence* in the **Into Variable** field.

5. Click **Apply**. The following text should appear in the window:

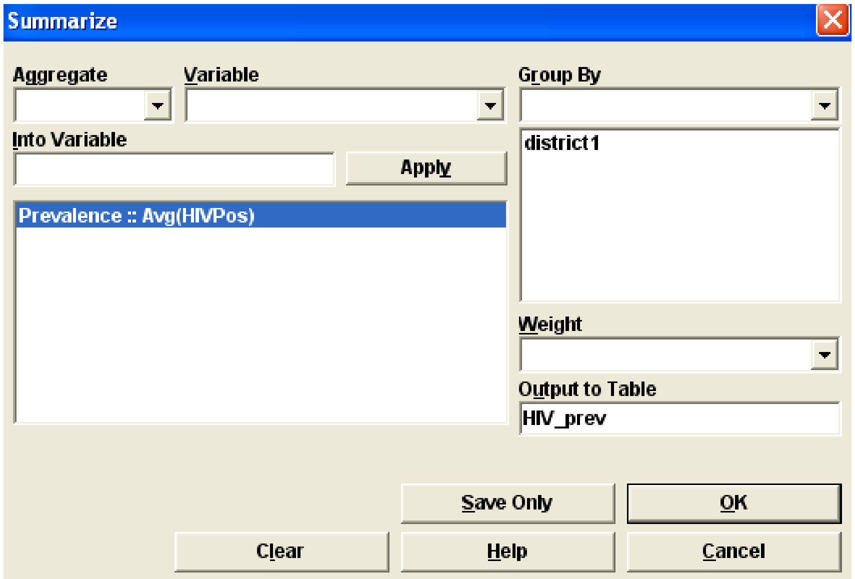
   Prevalence::Average(HIVPos)

To stratify your results by district:

6. In the **Group By** drop-down box, select *District1*.

Activity 8, Construct a Data Table for Epi Map, continued

7. In **Output To Table**, enter *HIV_Prev*. The **Summarize** dialogue box should appear as shown below.



Prevalence :: Avg(HIVPos)

Group By: district1

Weight:

Output to Table: HIV_prev

8. Click **OK**. You have properly constructed a data table to allow Epi Map to map your prevalence data.

Creating the Map

First, you need to have an Epi Info project (MDB) with a table containing at least two fields, including:

- the name of the geographical area you want to map
- the data you want to map.

Steps to create maps

Second, you need to have the necessary shape files with boundaries consistent with the geographical areas used in the Epi Info project file.

Follow these steps to create maps from Analysis:

1. **Read** the *C:\ANC_Suri\Analysis\ANCAll.mdb* project and select the *HIV_Prev* table. **List** the table to see the data for the geographical areas (*District1*) and the numeric field (*Prevalence*) you want to map.

2. Select the **Map** command from the command tree.

3. Check the **1 record per geographic entity** checkbox. Notice the **Aggregate Variable** field's value is changed to *Sum*, and disabled or greyed out.

4. From the **Geographic Variable** box on the left side of the dialog box, select *District1*. This field contains the names of districts.

5. From the **Data Variable** box on the left side of the dialog box, select *Prevalence.* This variable contains the prevalence data.

6. Select **Shapefile** in the middle of the dialog box. Select *C:\ANC_Suri\Maps\MH.SHP*.

7. From the **Geographic Variable** box below the Shapefile button, select *Name.* This field contains the names of geographical areas.

8. Verify that the geographic names are the same as the names from the Epi Info table by looking at the listing below the **Geographic Variable** box.

9. Select **OK**.

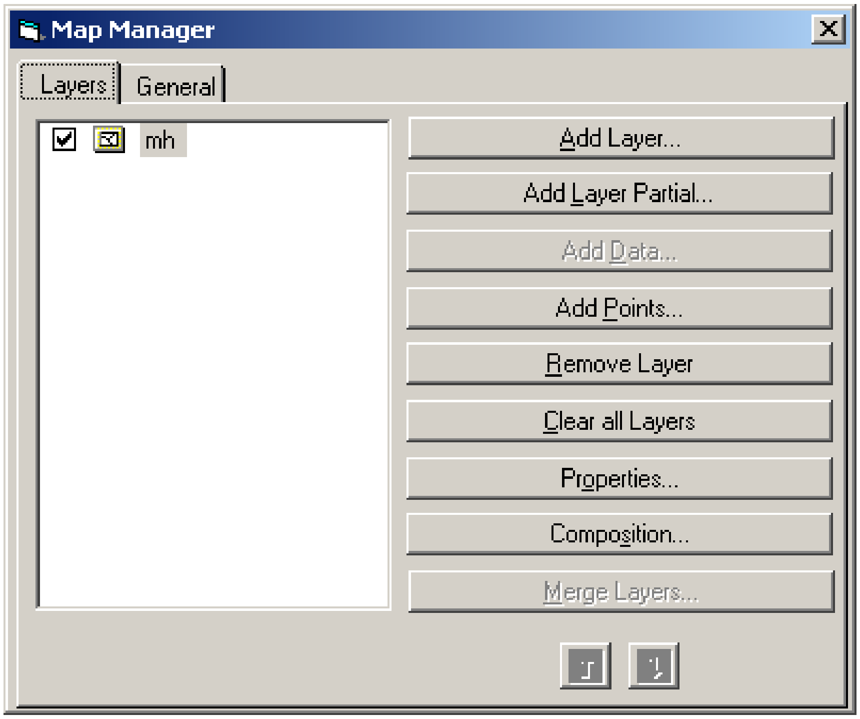
This will result in a map of prevalence rate by district.

Modifying Your Map

Using Map Manager

You can modify the colours, range and legend of your map using the **Map Manager** in **Epi Map**.

1. Select **Map Manager** from the **File** menu in **Epi Map**. Alternatively, click on the icon with three different coloured paper sheets on the upper left corner of the Epi Map output area. A dialog box with **Map Manager** on the title will appear.



The checkbox on the first line indicates that this map should be visible. The “mh” is the name of the shape file.

2. Select **Properties**. The Layer Properties for the MH dialog box is displayed.

3. Select the **Choropleth** tab.

Using Map Manager, continued

4. Select *SUM_prevalence* under **Numeric** field if it is not already selected.

   - Change the color ramp of the map by clicking on the white box next to the **Start**: prompt under Color Ramp. You will see a dialog box with color choices.
   - Select a new starting colour.
   - Repeat this process for the **End** colour. Epi Map will automatically determine the colour shades that should be used between the start value and the end value.
   - Click on **Reset Legend**. You will see that the colours have been changed to reflect your selection.
   - Check **Overlay** checkbox. Click **Apply**.

5. To display the names of districts, select the **Std Labels** tab.

   - Select **Name** in the Text field box.
   - Click **Apply** to display the labels.
   - Click **OK**.
   - Close **Map Manager**.

6. The map you have just created can be saved as a bitmap (BMP) graphic file and inserted in other documents. This can be done either by using the clipboard or by saving the image as a BMP file.

   - To copy the image to the clipboard and then paste it into a document, select **Edit** menu and **Copy Bitmap to clipboard**.
   - Open an application into which you want to paste the image (e.g., Microsoft® Word or PowerPoint), and then select Edit>>Paste.
   - From Epi Map, to save the image as a BMP (bitmap) file, select **File**>>**Save Bitmap File>>…**, specify a file location and name, and click **Save**.

Displaying Sites on Your Map

Epi Map can display point data (school, clinic or factory) on a map if X and Y co-ordinates are available. X and Y co-ordinates, also known as latitude and longitude, can be obtained by using a map or a global positioning system (GPS).

Steps to display point data

To display point data on your map, the steps are as follows:

1. In **Epi Map**, if the **Map Manager** is not visible, click on the **Map Manager** icon.

2. Select **Add Points**.

3. Select *C:\ANC_Suri\Maps\SiteNames* project, *SiteNames* table.

4. Highlight XCOORDINAT and YCOORDINAT for **X Field** and **Y Field**, respectively.

5. Select *Name* for **Point Label.**

6. Modify **Point Color**.

7. Change the **Point Size** to 5.

8. Select **OK**. You should see 19 sites with their names on the map. This map can be saved or copied to the clipboard and pasted elsewhere.

Creating the Map from Epi Map

The map can also be created from Epi Map using the same data and shape files. To create the same map from Epi Map, follow these steps:

Steps to create a map using Epi Map

1. Start **Epi Map**.

2. Start **Map Manager**.

3. Select **Add Layer…**

4. Select MH.SHP as your shape file. You should see the map on the screen.

5. Select **Add Data.**

6. Select *C:\ANC_Suri\Analysis\ANCAll.mdb* project, *HIV_Prev* table.

7. Under **Shape Fields Geographic Field**, select Name. This is the field that contains the names of geographical areas to be mapped.

8. Under *HIV_Prev Columns Geographic Field*, select *District1*. This is the field from *ANCAll: HIV_Prev* table that contains the names of geographical areas that will be matched with names from the shapefiles. This should already be highlighted.

9. Under ***HIV_Prev*** **Columns Render Field**, select *Prevalence*. This is the data field that will be displayed on the map.

10. Select **OK.** Epi Map will display a map with prevalence data. This map should be identical to the map that was created from Epi Info Analysis using the MAP command. This map can be modified using the Properties command, as discussed previously.

Notes

Exercise 10

Analysing Two or More Samples

Overview

What this exercise is about

In Exercise 9, we described the sample population characteristics using a variety of simple descriptive statistics. Most importantly, we calculated HIV prevalence among those sub-groups.

In this exercise, we want to determine whether the HIV prevalence in two populations differs from one to the other. For example, we will determine whether prevalence among urban participants, as compared to rural participants, is significantly higher, lower or not different. In addition, we will look at whether HIV prevalence differs significantly among those pregnant women aged less than 25, as compared to those aged 25 or greater.

To determine if any differences found in the rural/urban HIV prevalence are truly related to risk for contracting HIV and not related to possible differences in age distribution in the rural/urban areas, we will calculate age-standardised rural and urban HIV prevalence. We can then calculate the chi-square value for the age-adjusted rural/urban HIV prevalence percentages to see if a significant difference remains.

What you will learn

At the end of the exercise, you will be able to:

- calculate chi-square statistics to assess statistically significant differences between crude prevalence in two samples
- calculate chi-square statistics to assess statistically significant differences between age-standardised prevalence in two samples
- interpret and report results of chi-square.

Starting location

Analysis, C:\ANC_Suri\Analysis\ANCall.mdb:Analysis

Determining Statistical Differences

Often, policymakers want to know whether two populations differ with regard to their HIV prevalence in order to better target prevention efforts and resources. Populations to be compared may come from:

- two different sites, in which case we want to know whether one site has significantly lower or higher prevalence than the other
- two different sub-sets of the sample population; for example, from rural and urban women.

In the second example above, you demonstrated in Exercise 9 that the crude urban HIV prevalence among women sampled was 28.9% (878/3041), while HIV prevalence among rural women was higher at 33.4% (1141/3413). For the surveillance team, the important question is:

Is this difference significant?

HIV prevalence data table

To answer that question, we will determine whether HIV prevalence differs significantly among rural and urban attendees using a Yates-corrected chi-square (χ^2) test and its corresponding p-value. The two-by-two table of data showing Residence by HIV outcome for the calculation of the chi-square test is the same table that you set up in Exercise 9. It appears as follows:

Sample	Number of Persons HIV-positive	Number of Persons HIV-negative
Urban	$x^1 = 878$	$N2 - x^1 = 2163$
Rural	$x^2 = 1141$	$N2 - x^2 = 2272$

Calculating Yates-corrected chi-square

The Yates-corrected (χ^2) value can be calculated using the formula below:

$$(\chi^2) = \quad n(x^1\ (N^2 - x^2) - (N^2 - x^1)\ x^2 - .5n)^2$$

$$(x^1 + x^2)\ [N^2 - x^1) + N^2 - x^2)]\ [x^1 + (N^2 - x^1)]\ [x^2 + (N^2 - x^2)]$$

Fortunately, Epi Info can calculate the Yates-corrected chi-square value for us so we don’t have to do it by hand!

P-values

Using the chi-square value calculated by Epi Info, we can look up the p-values for the chi-square in a statistics book or use the p-value provided in Epi Info’s output. P-values equal to or less than 0.05 allow us to conclude that the prevalence rates for the two sample populations are significantly different. In general, it is useful to know that the larger the chi-square value, the smaller the p-value.

Calculating the chi-square statistic and P value

To calculate the chi-square statistic and p-value in Epi Info to determine whether HIV prevalence is statistically different in urban attendees as compared to rural attendees, follow the steps below:

1. **Read (Import)** the *C:\ANC_Suri\Analysis\ANCall.mdb* project file or type it into the project prompt box.

2. Select the **View All** to see the Analysis Table.

3. Click **OK**.

4. Select only those records where **Year= “2002**”. You should have 6 604 records in your sub-setted database.

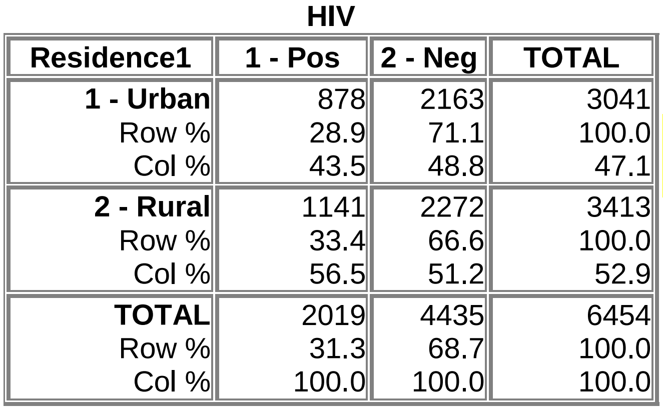
5. Select **Tables** from the command tree.

Calculating the chi-square statistic and P-value, continued

6. Select *Residence1* as the Exposure Variable from the drop-down list.
7. Select *HIV* as the Outcome Variable from the drop-down list.
8. Verify in your Settings that cases missing either *HIV* or *Residence1* are excluded from the analysis.
9. Click **OK**.

Results table

The results should appear as shown below:



HIV

Residence1	1 - Pos	2 - Neg	TOTAL
1 - Urban Row % Col %	878 28.9 43.5	2163 71.1 48.8	3041 100.0 47.1
2 - Rural Row % Col %	1141 33.4 56.5	2272 66.6 51.2	3413 100.0 52.9
TOTAL Row % Col %	2019 31.3 100.0	4435 68.7 100.0	6454 100.0 100.0

Single Table Analysis

	Point	95% Confidence Interval	
	Estimate	*Lower*	*Upper*
PARAMETERS: Odds-based			
Odds Ratio (cross product)	0.8083	0.7271	0.8985 (T)
Odds Ratio (MLE)	0.8083	0.7270	0.8985 (M)
		0.7260	0.8998 (F)
PARAMETERS: Risk-based			
Risk Ratio (RR)	0.8636	0.8027	0.9292 (T)
Risk Difference (RD%)	-4.5589	-6.8171	-2.3008 (T)
(T=Taylor series; C=Cornfield; M=Mid-P; F=Fisher Exact)			
STATISTICAL TESTS	Chi-square	1-tailed p	2-tailed p
Chi square - uncorrected	15.5480		0.0000816048
Chi square - Mantel-Haenszel	15.5456		0.0000817074
Chi square - corrected (Yates)	15.3367		0.0000911224
Mid-p exact		0.0000397783	
Fisher exact		0.0000442370	

Drawing conclusions from the results

The Yates-corrected chi-square (χ^2) is calculated above to be 15.34. This statistic has a p-value of 0.000091, which is less than the traditionally-used significance level of p<0.05.

We may therefore conclude that HIV prevalence in the populations represented by the two samples is indeed different. In this case, HIV prevalence among rural attendees [33.4% (1141/3413)] is significantly higher than HIV prevalence among urban attendees [28.9% (878/3041)].

Confidence intervals or P values

This process of deciding whether HIV prevalence rates are different may remind you of our work in Exercise 9, when we compared confidence intervals. Whether you determine differences using CIs or p-values, you will arrive at the same answer. Your choice to use and present p-values or CIs should be guided by your comfort in interpreting these statistics and your audience's ability to understand the information you are communicating.

Try it yourself!

Activity 1, Determine Significant Differences

Determine if there is a significant difference in HIV prevalence between women less than 25 years of age and women 25 years of age or greater. Be sure to exclude all records missing values for *AgeGroup*.

To do this, **Define** the variable *Age25* and **Assign** *Agegroup* to *Age25* using the following **If/Then** statements.

```
IF AgeGroup<=“20-24” THEN
    ASSIGN Age25=“<25”
ELSE
    ASSIGN Age25=“25+”
END

IF AgeGroup=(.) THEN
    ASSIGN Age25=(.)
END
```

Note that we used the second **If/Then** statement to make sure that missing values were properly coded. Without this statement, all cases missing an *AgeGroup* value would have been assigned a value of “25+” because of the **Else** command. Create a table of *AgeGroup* by *Age25*, with missing values included, to verify that you have recoded correctly.

You can then generate a table to determine whether women under 25 years have a higher or lower HIV prevalence than women age 25 years or greater.

For future age standardisation, use the **Write** command to create a separate table called *Subset* in C:\Suri\Analysis\ANCAll.mdb. The new table should only include the *Residence1*, *Age25*, and *HIV* variables.

Interpreting the results

Write a sentence describing HIV prevalence between women under 25 years and women aged 25 or greater. Indicate whether there is a significant difference between the HIV prevalence rates for these two populations of women.

Age Standardisation in a Two-Sample Comparison

While HIV prevalence by rural and urban participants can be compared as above, crude prevalence often needs to be adjusted to ensure that it is truly the risk difference between the rural or urban setting, as opposed to the woman's age, that makes HIV prevalence significantly higher in rural areas.

From our calculations above, we know that women aged 25 or greater are significantly more likely to have a higher HIV prevalence than women age 24 or less [38.1% (948/2488) and 27.1% (1065/3935) respectively, with a corresponding p-value less than 0.05 for the Chi square – corrected (Yates) value].

Accounting for age difference in populations

As such, if rural areas typically have older women, then we would expect higher prevalence in rural areas where women are younger. How can we figure out whether women in rural areas have a higher HIV prevalence simply because they are, on average, older than urban women?

From a frequency in Epi Info of *Age25* by *Residence1*, we can see that rural attendees, are in fact, older than urban attendees [i.e., 39.6% (1344/3397) of rural attendees are aged 25 or greater while 37.8% (1144/3026) of urban attendees are aged 25 or greater]. Thus, we might expect that the unadjusted HIV prevalence in the rural areas would actually be slightly lower if the age distributions in both the rural and urban areas were the same.

In our sample population, it is advisable to compare HIV prevalence estimates directly from rural and urban areas adjusted for age since the attendees in each category have different age distributions.

Adjusting for age

To remove the effect of different age distributions on rural/urban HIV prevalence, we use a process called direct adjustment or standardisation. By performing direct age adjustment, we can calculate the HIV prevalence for both urban and rural women as if they had the same age distributions instead of the age distributions they actually have. Using this age-adjusted prevalence, we can then determine whether the difference between urban and rural prevalence is still significant.

Adjusting for age: Step 1

To age-adjust in Epi Info, follow the three steps below:

Step 1 – Determine the percent distribution and total number of women by age group.

1. **Read** the *subset* table in *C:\ANC_Suri\Analysis\ANCall.mdb*.

2. **Define** the standard variable *One* and **Assign** *One*=1.

3. Calculate a frequency of *One* and write it out to a table called *N*.

4. Calculate a frequency of *Age25*, stratifying by *One* and writing it out to a table called *Age*.

5. **Read** the *Age* table you just created in *C:\ANC_Suri\Analysis\ANCall.mdb*.

6. Select the **Relate** command, show **All** views and choose **N**.

7. Select the **Build the Key Command**.

8. Using the Available Variables, relate the current and related tables by selecting *One*, clicking on **OK** and selecting *One* again, clicking again on **OK**.

9. Check the box labelled **Use Unmatched (ALL).**

10. Click **OK** to see the following text: RELATE N ONE :: ONE ALL

11. **Define** *P1*, the percent distribution of *Age25*.

12. **Assign** *P1*=COUNT/COUNT1.

Adjusting for age: Step 1, continued

13. **Write Replace** out the Table *T1* to *C:\ANC_Suri\Analysis\ANCall.mdb* containing only the variables *Age25* and *P1*.

Adjusting for age: Step 2

Step 2 – Determine the percentage distribution of women by age group within residence categories (urban/rural) to calculate the WEIGHT value.

1. Click **Read** to open *the C:\ANC_Suri\Analysis\ANCall.mdb* project file or type it into the project prompt box.

2. Select **All** to see the *Subset* Table.

3. Click **OK**.

4. Calculate a frequency of *Residence1*, outputting it to a Table you should name *Res*.

5. Calculate a frequency of *Age25*, stratifying by *Residence1* and outputting it to the Table *AgeRes*.

6. Read the *'C:\ANC_Suri\Analysis\ANCall.mdb' AgeRes* Table.

7. Select **Relate** and the Table *Res*, **Build**ing the key using *Residence1* to *Residence1*.

8. **Define** *P2* and **Assign** *P2=Count/Count1*.

9. Select **Relate** and the Table *T1* and then **Build** the key using *Age25* to *Age25*.

10. **Define** *Weight*.

11. **Assign** *Weight=P1/P2*.

12. **Write** the variables *Residence 1*, *Age25* and *WEIGHT* out to the table *T2* in the *C:\ANC_Suri\Analysis\ANCall.mdb* Project.

Adjusting for age:
Step 3

Step 3 – Calculate the standardised HIV prevalence by residence adjusting for age using the weights.

1. Click **Read** to open the *C:\ANC_Suri\Analysis\ANCall.mdb* project file or type it into the project prompt box.

2. Select **All** and then select the *Subset* Table.

3. Click **OK**.

4. Select **Relate** and the Table *T2*, then **Build** the key using *Age25* to *Age25* AND *Residence1* to *Residence1*.

5. Calculate the HIV prevalence by Residence using the *Weight* variable as the Weight.

Try it yourself!

Activity 2, Describe HIV Prevalence Findings

Describe your findings of HIV prevalence by residence using the age-adjusted weights. Compare the age-adjusted rural/urban HIV prevalence to the crude HIV rural/urban prevalence you calculated in Exercise 9. Did the prevalence values change accordingly, based on our knowledge that HIV prevalence was higher among the older age groups and higher in rural residents?

Exercise 11

Comparing Three or More Samples (Time Trends)

Overview

What this exercise is about

In Exercise 11, you investigated the chi-square test for significant differences between two populations: rural vs. urban. In addition, you looked at HIV prevalence of women < aged 25 and 25 years or older.

Also of interest is whether HIV prevalence is increasing or decreasing over time in Suri, and for these sub-groups in particular. In this exercise, we will determine whether or not changes (i.e., increases or decreases) have occurred in annual HIV prevalence from 2000 to 2002.

What you will learn

At the end of the exercise, you will be able to:

- conduct a chi-square test for linear trends
- interpret and report results
- construct and interpret line graphs.

Starting location

Analysis.

Resources

Exercise.

Determining Statistical Difference Over Time

A common method for determining whether or not changes (increases or decreases) have occurred in annual HIV prevalence over time is to calculate chi-square tests for linear trends. The test statistic is also known as the chi-square of slope since it calculates the probability that the change in prevalence (or slope of the line over time) is changing.

ANC 2000-2002 example

Consider the following example of data collected from ANC surveys conducted from 2000-2002 that are used in the calculation of the chi-square for slope. Fill in the table below with Suri data.

Year	Number of Persons Tested	Number HIV-positive	Number HIV-negative	Estimated HIV-Prevalence
2000				
2001				
2002				

From the table, we might conclude that the HIV prevalence in Suri is declining over time. Both for the population and for the policymakers, this decline in estimated HIV prevalence among the population of pregnant women is a positive sign in the fight against HIV infection.

Did the decline really occur?

The question for epidemiologists, however, is whether the decline over time in HIV difference is real. To determine the answer to this question, we first have to sub-set the data to include only those sites that have participated in the survey each year.

Creating a sub-set

At a minimum, three data points over time should be available to analyse trends.

1. Click **Read** to open the *C:\ANC_Suri\Analysis\ANCall.mdb* project file or type it into the project prompt box.

2. Select **All** to see the *Analysis* Table.

3. Click **OK**.

4. Create a **Table** distribution of *SiteName* by *Year*. Identify those sites that do not have three years of data. Write the names of the sites in the spaces provided.

   ____

   ____

   ____

   ____

5. Sub-set your data using the **Select** command to exclude these sites.

6. Next, use the **Tables** function to calculate HIV prevalence by year in the sub-setted database. The following table should appear below:

Year	1 - Pos	2 - Neg	TOTAL
2000	1573	3342	4915
Row %	32.0	68.0	100.0
Col %	31.4	30.5	30.8
2001	1709	3742	5451
Row %	31.4	68.6	100.0
Col %	34.1	34.1	34.1
2002	1732	3880	5612
Row %	30.9	69.1	100.0
Col %	34.5	35.4	35.1
TOTAL	5014	10964	15978
Row %	31.4	68.6	100.0
Col %	100.0	100.0	100.0

Creating a sub-set, continued

Fill in the table below.

Year	Number of Persons Tested	Number HIV-positive	Number HIV-negative	Estimated HIV Prevalence
2000				
2001				
2002				

To investigate whether this decline is statistically significant, we will use StatCalc in Epi Info. Once again, we are interested in the chi-square statistic and the associated p-value.

Statistical significance can also be calculated using logistic regression methods in Epi Info Analysis, although the process of creating analysis variables is more complex. Note that the chi-square that you get in the tables function for more than two variables is not the chi-square test for trend value.

Using StatCalc

To calculate the chi-square of slope statistic and p-value:

1. Run **StatCalc** from the Epi Info Utilities menu.

2. Select the **Chi Square for trend** option.

3. Type *2000* into the Exposure Score category. Press **Enter**.

4. Type in the number of HIV-positive persons in the year 2000 into the cases column. Press **Enter**.

5. Type in the number of HIV-negative persons in the year 2000 into the controls column. Press **Enter**.

6. Continue typing in data for the years 2001 and 2002.

7. Press the function key **F4** to calculate the chi-square of slope.

Interpreting results

The chi-square test for linear trend has a value of 1.58 and an associated p-value of 0.21, which is not statistically significant at the 0.05 level. It can therefore be concluded that the decline in HIV prevalence among clinic attendees from 2000-2002 was not significant.

Try it yourself!

Activity 1, Calculate Suri HIV Prevalence Over Time

Calculate HIV prevalence over time by Residency in Suri. Remember to exclude those records that are missing residence.

- Write a couple of sentences that summarise whether rural and urban trends are increasing or decreasing and whether this difference is statistically significant.
- Generate a bar chart of your results by year and sub-category.

Activity 2, Determine if HIV Prevalence Is Increasing

Calculate HIV prevalence over time (Years 2000 – 2002) to determine if HIV prevalence in women <25 years is increasing. Compare this trend to that of women aged 25 years or older over time.

- Generate a bar chart of your results by year and sub-category.
- Describe whether the increase or decrease is statistically significant.

Notes

Exercise 12
Developing a National Report

Overview

What this exercise is about

Analysis of the 2000-2002 Suri surveillance data is complete. The surveillance team and stakeholders have agreed that the analysis accurately captures HIV prevalence among the population of pregnant women aged 12–49 sampled in our survey. Most importantly, with this knowledge, we can create a national report. This report is intended to inform other public health scientists and policymakers in Suri, but it may be disseminated and used by the national and international press to describe HIV prevalence in the country.

Using Microsoft Word or PowerPoint, you will create a national or regional report describing the results of your work throughout the previous week. This result will communicate your findings to a variety of different audiences.

What you will learn

At the end of the exercise, you will be able to:

- include data and graphs from Epi Info analysis output in Microsoft Word or PowerPoint documents
- list the basic format of a national report on HIV sentinel surveillance among pregnant women.

Starting location

Analysis.

Resources

Previous exercises.

Using Epi Info with Microsoft Word and PowerPoint

National reports are an important means for communicating your results to stakeholders and the public. Without a dissemination component, the work of collecting, cleaning and analysing your data is of little use.

Epi Info provides great tools for data entry, data management and analysis. The Epi Info Report Writer tool also provides the capacity to produce standardised reports. While these reports are generally easy to read and produce, they often lack the professional document style required for many publications. As a result, we can also produce reports in other applications, such as Microsoft Word or PowerPoint. Those tools obviously don't provide the ability to analyse our data, so we must be able to take Epi Info analysis output and copy it to Microsoft Word or PowerPoint.

Copying Epi Info Text and Table Output to Microsoft Word or PowerPoint

Epi Info Analysis produces output from analysis in the Display Window using the HTML file format. Copy output, such as frequency tables or lists, from the Display Window, then paste into other applications, such as Microsoft Word or PowerPoint in Windows.

Copying Epi Info text

Use the standard methods for copying. Highlight the section you want to copy, then:

- click Ctrl+C to copy and Ctrl+V to paste, or
- right-click and select **Copy**, then right-click and select **Paste.**

To select, copy and paste output from Epi Info to Microsoft Word or PowerPoint for a table of *AgeGroup*:

1. Open Microsoft Word or PowerPoint and place your cursor where you want to copy the data from Epi Info.

2. Open or switch to the Analysis application in Epi Info.

3. Click **Read** to open the *C:\ANC_Suri\Analysis\ANCall.mdb* project file or type it into the project prompt box.

Copying Epi Info text, continued

4. Select the **View All** to see the *Analysis* Table.

5. Click **OK**.

6. Generate a frequency of *AgeGroup* in Epi Info for Suri, year 2002.

7. Left-click on the corner of the table in the Display Window and drag the cursor over the table to highlight it.

8. While it is still highlighted, press Ctrl+C or right-click and select **Copy**. Your frequency table should now be copied to the Microsoft Windows Clipboard.

9. You can now go to Microsoft Word and paste the table. You can either:

   - use the paste command under the Edit menu
   - right-click and select paste
   - press Ctrl+V.

Try it yourself!

Activity 1, Generate an HIV Prevalence Table

Generate a table of HIV prevalence for each age group. Copy and paste the results from Epi Info to either Microsoft Word or PowerPoint.

Copying Epi Info Graphs and Charts to Microsoft Word or PowerPoint

Epi Info Analysis produces graphs and charts in the Analysis Display Window that can be saved as .jpg files. The graphs and charts in EpiGraph can be copied by selecting the Menu>>Edit>>Copy to Clipboard commands.

Once the graph has been created and is displayed in the Display Window, you can highlight the graph and press Ctrl+C, as you did when copying the table output.

Steps to select, copy and paste graphs

To select, copy and paste a graph from Epi Info to Microsoft Word or PowerPoint:

1. Open Microsoft Word or PowerPoint and place your cursor where you want to copy the graph from Epi Info.

2. Open or switch to the Analysis application in Epi Info.

3. Click **Read** to open the *C:\ANC_Suri\Analysis\ANCall.mdb* project file or type it into the project prompt box.

4. Select the **View All** to see the *Analysis* Table.

5. Click **OK**.

6. Generate a graph of HIV prevalence by *AgeGroup* in Epi Info for Suri, year 2002.

7. Select **Edit** from the EpiGraph Menu.

8. Click on **Copy to Clipboard**.

9. Your graph should now be copied to the Microsoft Windows clipboard.

10. You can now go to Microsoft Word and paste the table, using either the Paste command under Edit in the Menu file, right-clicking and selecting paste or pressing Ctrl+V.

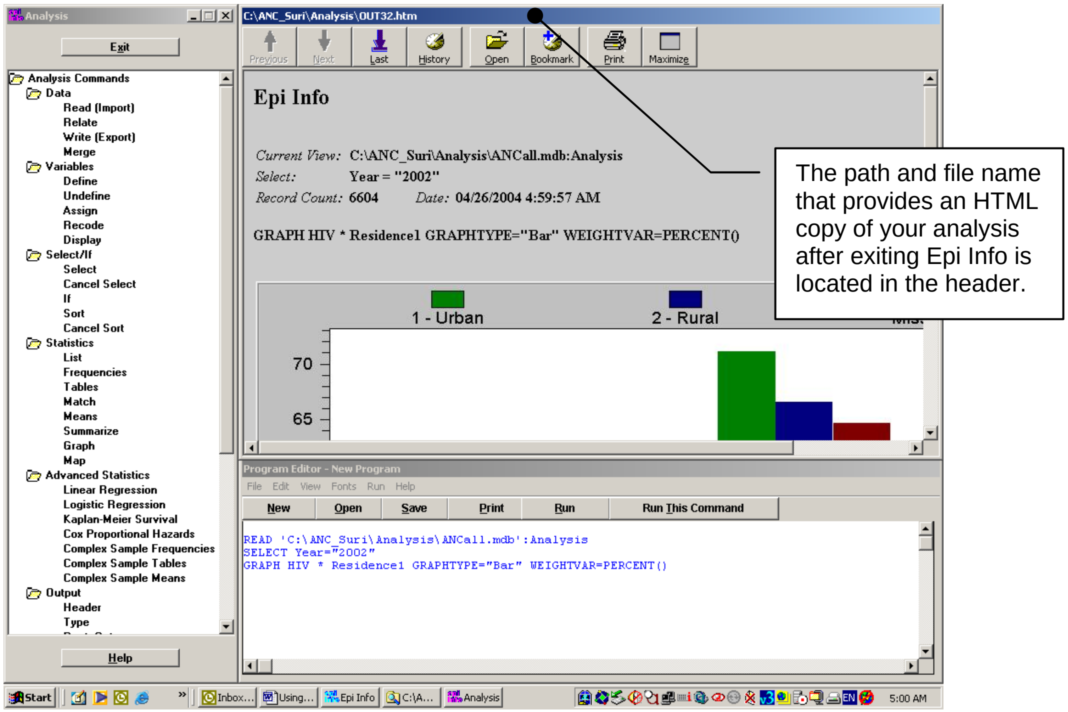
Try it yourself!

Activity 2, Generate an HIV Prevalence Graph

Generate a graph of HIV prevalence by residence. Copy and paste the results from Epi Info in either EpiGraph or the Display Window to either Microsoft Word or PowerPoint.

Accessing Epi Info Analysis HTML Output

The results of each Analysis session displayed in the Display Window are saved as an HTML file in the Project's File Folder as Out#.htm, where # is the number of Analysis sessions that have been opened. Any graphs or maps are saved separately as Out#_#.jpg, where the number after the underscore is the number of graphs that have been produced in the session.



Epi Info

Current View: C:\ANC_Suri\Analysis\ANCall.mdb:Analysis
Select: Year = "2002"
Record Count: 6604 *Date:* 04/26/2004 4:59:57 AM

GRAPH HIV * Residence1 GRAPHTYPE="Bar" WEIGHTVAR=PERCENT()

```
READ 'C:\ANC_Suri\Analysis\ANCall.mdb':Analysis
SELECT Year="2002"
GRAPH HIV * Residence1 GRAPHTYPE="Bar" WEIGHTVAR=PERCENT()
```

The path and file name that provides an HTML copy of your analysis after exiting Epi Info is located in the header.

Viewing HTML output

Epi Info stores the Display Window output as a file; the filename can be found at the top of the Display Window file. Results from these HTML files can be viewed in any internet browser or by opening the file in Microsoft Word.

Try it yourself!

Activity 3, Find the File in Windows Explorer

Find the file in Windows Explorer or the analysis that you just generated when creating the table and graph above in Epi Info, and open it using your browser.

Components of a National Report

National reports should provide the most important information about the HIV ANC survey in an easy-to-read format. Typically, we include the following sections in a national report:

What national reports typically include

Table of Contents

- Titles of major section areas

List of Tables and Figures

- Titles of tables and figures

Executive Summary

- One page summary of the report

Background

- Information about the country
  - Population size
  - Urban/rural distribution

- History of the epidemic in the country
  - When the epidemic started
  - When surveillance started
  - Types of surveillance surveys in the country (past and present)
  - History of ANC surveillance

Objectives

- Goals and objectives of the ANC surveillance survey

What national reports typically include, continued

Methods

- Study design
  - Anonymous and unlinked
- Study population
  - Patient inclusion criteria
- Sentinel clinic sites
  - Site inclusion criteria
- Sample size
  - Overall
  - By clinic
- Data collection tools
  - Description of forms
- Laboratory methods
  - Method of blood collection
  - Preparation of specimens
  - Testing laboratories
  - Quality assurance procedures
  - Tests
  - Testing algorithm
  - Storage and transport of specimens
- Data management
  - Flow and storage of data collection forms
- Training and supervision
  - Regionally
  - Clinic sites
  - Laboratory

What national reports typically include, continued

Results

- Characteristics of the study population
- HIV and syphilis seroprevalence
- HIV seroprevalence by
  - Age group
  - Marital status
  - Education
  - Residence
  - Occupation
  - Gravida
  - Region
  - Clinic site
  - Comparison of two years of prevalence data
- Trends in HIV prevalence over time
  - Overall prevalence
  - Age group
  - Education
  - Residence
  - Region
  - Clinic site
- Populations of interest
  - Young women (aged 15-24)
    - HIV seroprevalence
    - Trends over time
  - Rural/Urban attendees

Discussion

- Discussion of findings and data limitations
- Possible reasons for findings/factors influencing the results
- Differences from previous years' ANC surveys or other seroprevalence surveys and possible reasons for those differences
- Strengths and limitations of conducting surveillance in antenatal clinics (country-specific)
- Generalisation of findings to non-pregnant women and to men (epidemiologic projections, if done)

What national reports typically include, continued

Conclusions and Recommendations

- Summary of findings
- Recommendations for the next round of ANC surveillance
- Recommendations for use of the data for policy, prevention and care

In the results section, you should be able to use your experiences with entering, managing and analysing data to produce data for decision-making.

Activity 4, Produce the Suri National Report

For the final activity, produce a national report for Suri, focusing on the **Objectives**, **Methods**, **Results**, **Discussion** and **Conclusions and Recommendations** Sections. Be sure to include results that describe your special populations, such as women by age group and the rural/urban population.

Once completed, you can begin to the think about next year! Congratulations on your first successful year as a new member of the Suri HIV Surveillance Team.

Notes

Appendix A

Hardcopy forms begin on the next page.

To access the forms using Adobe Acrobat reader:

1. Right-click the Start menu button.

2. Click on **Explore.**

3. Type *C:\ANC_Suri\Documentation\Appendix A.pdf* in the address navigation bar to open the Adobe Acrobat PDF file to see data collection forms.

4. Print the forms.

Appendix B,
HIV Surveillance Data Collection Form for ANCs—WHO Recommended

Ministry of Health
HIV Surveillance Data Collection Form for Antenatal Clinics

Site: ____ **District:** ____

Demographic Information:

Survey ID Code: □

Date of patient visit (dd/mm/yyyy):_____/_____/________ **Age (in years): _______**

Residence: □ **Urban** □ **Rural** □ **Missing**

Highest level of school attended: □ **None** □ **Primary** □ **Secondary** □ **Higher** □ **Missing**

Occupation (primary): (optional)

□ **Business** □ **Housewife** □ **Not employed**
□ **Domestic help** □ **Police/Military** □ **Other**
□ **Student** □ **Laborer** □ **Missing**
□ **Farmer** □ **Professional**

Total number of pregnancies, including this pregnancy: □

Total number of live births: □

Test Result Information:

HIV	**Screening (Initial Test) Date:** _____/_____/________ **dd mm yyyy**	□ **Positive**	□ **Negative**
	Confirmatory Test Date: _____/_____/________ **dd mm yyyy**	□ **Positive**	□ **Negative**
Syphilis	**RPR date:** _____/_____/________ **dd mm yyyy**	□ **Positive**	□ **Negative**

Appendix C,
Suri Surveillance Data Collection Form for ANC (YR.2001)

Suri Ministry of Health
Antenatal Care Clinic Data Collection Form
Round 2 - Year 2001

Survey Information

Survey Site Name:

Survey ID Number:

Demographic Information

Date of Patient Visit (dd/mm/yyyy): ____ / ____ / ________ **Age (in years):**

Residence: □ **Rural** □ **Urban**

Highest level of school attended: □ **None** □ **Primary** □ **Secondary** □ **Higher**

Marital Status:
□ **Single, never married** □ **Divorced**
□ **Married** □ **Widowed**

Occupation (primary):
□ **Business** □ **Domestic help**
□ **Police/military** □ **Laborer**
□ **Security guard** □ **Professional**
□ **Student** □ **Not employed**
□ **Farmer** □ **Other**
□ **Housewife**

Total number of pregnancies, including this pregnancy:

Total number of live births:

Laboratory Test Result Information

HIV Test □ **Positive** □ **Negative** **Date (dd/mm/yyyy):** ____ / ____ / ________

RPR Syphilis Test □ **Positive** □ **Negative**

Appendix D,
Suri Surveillance Data Collection Form for ANC (YR.2002))

Suri Ministry of Health

HIV Surveillance Data Collection Form for Antenatal Care Clinics
Round 3 - Year 2002

Survey Information

Survey Site Name: □

Survey ID Number: □

Demographic Information

Date of Patient Visit (dd/mm/yyyy): ____ / ____ / ________ **Age (in years):** □

Residence: □ **Urban** □ **Rural**

Highest level of school attended: □ **None** □ **Primary** □ **Secondary** □ **Higher**

Marital Status:

□ **Single, never married**	□ **Divorced**
□ **Married**	□ **Widowed**

Occupation (primary):

□ **Business**	□ **Domestic help**
□ **Police/military**	□ **Laborer**
□ **Security guard**	□ **Professional**
□ **Student**	□ **Not employed**
□ **Farmer**	□ **Other**
□ **Housewife**	

Total number of pregnancies, including this pregnancy: □

Total number of live births: □

Laboratory Test Result Information

HIV Test	□ **Positive**	□ **Negative**	**Date (dd/mm/yyyy):** ____ / ____ / ________
RPR Syphilis Test	□ **Positive**	□ **Negative**	**Date (dd/mm/yyyy):** ____ / ____ / ________
TPHA Syphilis Test	□ **Positive**	□ **Negative**	**Date (dd/mm/yyyy):** ____ / ____ / ________

Appendix E,
Data Dictionary for the Suri ANC Survey

Entity	Variable Prompt	Type	Size	Field Name	Code Table Values	Comments	Version Control
Location	Site Name*	Text	2	sit_num	"01"–Banket "02"–Chema "03"–Chickry "04"–Cholai "05"–Danu "06"–Goma "07"–Gwana "08"–Hidim "09"–Istan "10"–Kabi "11"–Karanda "12"–Loma "13"–Maka "14"–Mindi "15"–Mura "16"–Mustubini "17"–Nabo "18"–Nkula "19"–Tapanda	See check code section; Table Name: codeSit_num	"02-Added Nov 1999
	District	Text	1	district	"1"–Tibul "2"- Mandor "3"–Rikura "4"–Yemenia "5"–Insa "6"–Karafam "7"–Ashra	Read Only variable populated by Site Name Code Table. See Table Name: codeSit_num	Added Nov 1999
	?????	Text	3	Region	"MVG"–Mavinga "MAS"–Masana "HAR"–Hatar "MAN"–Malange	Read Only variable populated by Site Name Code Table. See Table Name: codeSit_num	Added Nov 1999

Appendix E,
Data Dictionary for the Suri ANC Survey

Entity	Variable Prompt	Type	Size	Field Name	Code Table Values	Comments	Version Control
Patient Identifiers	Unique Form ID	Text	6	pt_key	Calculated field	Read Only variable. Unique per client during one year; repeated over years.	Added Nov 1999
	Survey ID*	Text	3	id_num	Unique sequential code at site	Text field requires use of leading zeros	Added Nov 1999
Demographic	Patient Visit Date	Date	10	vst_date	dd-mm-yyyy	See check code section	Added Nov 1999
	Age (in years)	Num	3	**???**	12-49 998–Missing 999–Unknown	See check code section Range should be set from (12-49).	Added Nov 1999
	Residence	Text	2	residence	"1"-Urban "2"-Rural "98"–Missing	Table name: codeResidence	Added Nov 1999
	Highest School Level	Text	2	educ_leva	"1"–None "2"–Primary "3"–Secondary "4"–Higher "98"–Missing	See check code section; Table Name: codeEduc_leva	Added Nov 1999
	Marital Status	Text	2	mar_stat	"1"–Single "2"–Married "3"–Divorced "4"–Widowed "98"-Missing	Table name: codeMar_stat	Added Nov 1999

Entity	Variable Prompt	Type	Size	Field Name	Code Table Values	Comments	Version Control
	Occupation	Text	**?**	occup	"1"–Business "2"-Police/Military "3"-Security Guard "4"–Student "5"–Farmer "6"–Housewife "7"-Domestic Help "8"–Laborer "9"–Professional "10"-Not Employed "11"-Other "998"–Missing	Table name: codeOccup	Added Nov 1999
	Total pregnancies	Num	3	grav	1-15 998–Missing	See check code section	Added Nov 1999
	Total Live Births	**???**	3	par	1-15 998–Missing	See check code section	Added Nov 1999
Lab	HIV-1 Result	Text	2	HIV_res	"1"–Positive "2"–Negative "98"–Missing	See check code section; Table name: HIV_res	Added Nov 1999
	HIV Test Date	Date	10	hiv_date	dd-mm-yyyy	See check code section	Added Nov 1999
	Syphilis Result (RPR)	Text	2	RPR_res	**??**	See check code section; Table name: RPR_res	Added Nov 1999

Appendix F,
Check Code and Documentation for the Suri HIV Surveillance System

Variable	Checkcode	Action Trigger	Reference Variables
ID_Num	*This Check Code Assigns a Unique form ID (called *Pt_key) to the record based on the district, site and the ID_num). The Unique for ID must be 6 characters in size. ASSIGN Pt_key=SUBSTRING(District,1,1)& Substring(sit num,1,2)& Substring(id num,1,3)	After ID_Num	Pt_key District Sit_num
Vst_dt	*The Visit Date (vst_date) variable requires the date for patient visit to be within the allowable time period of the ANC sentinel surveillance round. In this example, the start date is 01/01/2002 with an end date of 31/12/2002. IF Vst_date<01/01/2002 OR Vst_date>12/31/2002 THEN DIALOG "Please enter a date between 01/01/2002 and 31/12/2002" TITLETEXT="Invalid Date Range" CLEAR Vst_date GOTO Vst_date END	After Vst_dt	
Age		After Age	None

Educ_leva	*Education level should not negatively correspond *to age. For example, a woman who has an age of 13 should not have completed higher education.
	
	IF Educ_leva="4" and Age<16 THEN
	DIALOG "Please insure that age is entered correctly. Age and education level are not consistent."
	TITLETEXT="Possible Data-entry Error"
	END
		After Educ_leva	Age

Par			

Variable	Checkcode	Action Trigger	Reference Variables

HIV_res	*If no test was done, then HIV test date is hidden
	IF HIV_res="98" THEN
	      HIDE Hiv_Date
	END
		After HIV_res	Hiv_date

TPHA_res			

RPR_res	* If no test was done, then RPR test date is hidden
	IF RPR_res="98" THEN
	HIDE Rpr_Date
	END
		After RPR_res	Rpr_date

Hiv_date	*Check for consistency of dates IF Hiv_Date<01/01/2002 OR Hiv_Date>12/31/2002 THEN DIALOG "Please input a date between 01/01/2002 and 12/31/2002." TITLETEXT="Invalid Date Range" CLEAR Hiv_Date GOTO Hiv_Date END	After Hiv_date	
Rpr_date			
Tpha_date			

Appendix G

Hardcopy forms for Appendix G follow.

For those with access to the free Adobe Acrobat reader,

1. Right-click the Start menu button.

2. Click on **Explore.**

3. Type: C:\ANC_Suri\Documentation\Appendix G.pdf in the address navigation bar to open the Adobe Acrobat PDF file to see the 5 data collection forms.

4. Print the forms.

Appendix G,
Banket HIV ANC Surveillance Data Collection Forms to be Entered

Appendix H.1,
HIV Surveillance Data Entry Audit Log – 2002

Column A	**Column B**	**Column C**	**Column D**	**Column E**	**Column F**	**Column G**	**Column H**	**Column I**
Date	**Survey Site Name (if using form)**	**Survey ID Number (if using form)**	**Unique ID Number (if electronic entry)**	**Variable name and current value**	**Description of problem**	**Resolution description**	**Date of final resolution**	**Initials of Data Mgr**

Appendix H.1,
HIV Surveillance Data Entry Audit Log – 2002

Column A	Column B	Column C	Column D	Column E	Column F	Column G	Column H	Column I
Date	**Survey Site Name (if using form)**	**Survey ID Number (if using form)**	**Unique ID Number (if electronic entry)**	**Variable name and current value**	**Description of anomaly**	**Resolution description**	**Date of final resolution**	**Initials of Data Mgr**

Appendix H.1,
HIV Surveillance Data Entry Audit Log – 2002

Column A	Column B	Column C	Column D	Column E	Column F	Column G	Column H	Column I
Date	Survey Site Name (if using form)	Survey ID Number (if using form)	Unique ID Number (if electronic entry)	Variable name and current value	Description of anomaly	Resolution description	Date of final resolution	Initials of Data Mgr

Notes

Appendix H.2,
HIV Surveillance Data Entry Audit Log – 2001

Column A	Column B	Column C	Column D	Column E	Column F	Column G	Column H	Column I
Date	Survey Site Name (if using form)	Survey ID Number (if using form)	Unique ID Number (if electronic entry)	Variable name and current value	Description of anomaly	Resolution description	Date of final resolution	Initials of Data Mgr

Notes

Appendix I

Hardcopy forms for Appendix I follow.

For those with access to the free Adobe Acrobat reader,

1. Right click the Start menu button.
2. Click on **Explore.**
3. Type: C:\ANC_Suri\Documentation\Appendix I.pdf in the address navigation bar to open the Adobe Acrobat PDF file to see the data collection forms.
4. Print the forms.

Notes