

PARTICIPANT WORKBOOK



Creating an Analysis Plan

Created: 2013



Creating an Analysis Plan. Atlanta, GA: Centers for Disease Control and Prevention (CDC), 2013.

Creating an Analysis Plan

Table Of Contents

INTRODUCTION.....	3
LEARNING OBJECTIVES.....	3
ESTIMATED COMPLETION TIME.....	4
TARGET AUDIENCE.....	4
PREWORK AND PREREQUISITES	4
ABOUT THIS WORKBOOK AND THE ACTIVITY WORKBOOK	4
ICON GLOSSARY	4
ACKNOWLEDGEMENTS	5
SECTION 1: OVERVIEW OF DATA ANALYSIS.....	6
STEPS TO COMPLETE BEFORE ANALYZING DATA.....	6
OVERVIEW OF STEPS IN ANALYZING NCD DATA	6
TYPES OF STATISTICAL DATA.....	8
SECTION 2: ANALYSIS PLAN	11
OVERVIEW	11
RESEARCH QUESTIONS AND/OR HYPOTHESES	11
DATASET(S) TO BE USED	11
INCLUSION/EXCLUSION CRITERIA	13
VARIABLES TO BE USED IN THE MAIN ANALYSIS	13
STATISTICAL METHODS AND SOFTWARE	15
KEY POINTS TO REMEMBER.....	16
OVERVIEW OF PREPARING TABLE SHELLS	23
TYPES OF TABLE SHELLS	24
KEY POINTS TO REMEMBER.....	32
RESOURCES	38
APPENDICES.....	40
APPENDIX A	41
APPENDIX B	42

Introduction

The **Creating an Analysis Plan** training module is one of three modules that will provide you with the skills needed to analyze and interpret quantitative¹ noncommunicable disease (NCD) data. When you apply these quantitative analysis skills, you will turn data into information that can be used to make informed decisions on public health program and policy recommendations.



An analysis plan helps you think through the data you will collect, what you will use it for, and how you will analyze it. Creating an analysis plan is an important way to ensure that you collect all the data you need and that you use all the data you collect.

Analysis planning can be an invaluable investment of time. It can help you select the most appropriate research methods and statistical tools. It will ensure that the way you collect your data and structure your database will help you get reliable analytic results.

LEARNING OBJECTIVES

Given information about a noncommunicable (NCD) health problem and a request for health-related information, you will be able to create an analysis plan that includes the following:

- Research question(s) and/or hypotheses,
- Dataset(s) to be used,
- Inclusion/exclusion criteria,
- Variables to be used in the main analysis,
- Statistical methods and software to be used, and,
- Table shells to prepare for:
 - Univariable analysis,
 - Bivariable analysis,

¹ Collecting, analyzing, and reporting qualitative data is a valuable epidemiologic skill that requires careful consideration but will not be covered in this module.

- Calculating measures of association, and,
- Assessing for confounding and effect measure modification.

ESTIMATED COMPLETION TIME

The workbook should take between 6 and 7 hours to complete.

TARGET AUDIENCE

The workbook is designed for FETP residents who specialize in NCDs; however, you can also complete the module if you are working in the infectious disease area.

PRE-WORK AND PREREQUISITES

Before participating in this training module, you must complete training in:

- Basic epidemiology and surveillance
- Basic analysis

ABOUT THIS WORKBOOK AND THE ACTIVITY WORKBOOK

The format of the **Participant Workbook** consists of 3 sections. You will read information about creating an analysis plan and complete 2 exercises to practice the skills and knowledge learned. At the end of the training module you will access the **Activity Workbook** and complete a skill assessment which combines all skills taught.

ICON GLOSSARY

The following icons are used in this workbook:

Image Type	Image Meaning
 Activity Icon	Activity, exercise, assessment or case study that you will complete
 Stop Icon	Stop and consult with your facilitator/mentor for further instruction
 Tip Icon	Supplemental information, or key idea to note and remember

ACKNOWLEDGEMENTS

Many thanks to the following colleagues from the Centers for Disease Control and Prevention for:

1) Providing detailed feedback and guidance:

- Lina Balluz, ScD, MPH, Office of Surveillance, Epidemiology and Laboratory, Division of Behavioral Surveillance
- Richard Dicker, MD, MS, Center for Global Health, Division of Global Health Protection
- Antonio Neri, MD, MPH, National Center for Chronic Disease Prevention and Health Promotion, Division of Cancer Prevention and Control
- Mona Saraiya, MD, MPH, National Center for Chronic Disease Prevention and Health Promotion, Division of Cancer Prevention and Control

2) Developing the hypertension case study for the Practice Exercises:

- Fleetwood Louston, PhD, FNP, National Center for Chronic Disease Prevention and Health Promotion, Division of Heart Disease and Stroke Prevention
- Andrea Neiman, MPH, PhD, National Center for Chronic Disease Prevention and Health Promotion, Division of Heart Disease and Stroke Prevention
- Cathleen Gillespie, MS, National Center for Chronic Disease Prevention and Health Promotion, Division of Heart Disease and Stroke Prevention
- Edward Gregg, PhD, National Center for Chronic Disease Prevention and Health Promotion, Division of Diabetes Translation

Some of the content of this module was taken from a training manual developed by the CDC's Division of Epidemiology and Surveillance Capacity Development: *Advanced Management and Analysis of Data Using Epi Info for Windows: Risk Factors for Sexually Transmitted Infections in Kuwadzana, Zimbabwe; 2006.*

Section 1: Overview of Data Analysis

STEPS TO COMPLETE BEFORE ANALYZING DATA

There are several steps you must complete before you analyze data. For this training, these steps have been divided into two modules – Create an Analysis Plan and Manage Data. The main tasks are as follows:

1. Create an analysis plan

- Identify research questions and/or hypotheses.
- Select and access a dataset.
- List inclusion/exclusion criteria.
- Review the data to determine the variables to be used in the main analysis.
- Select the appropriate statistical methods and software.
- Create table shells.

2. Manage the data

- Create a data dictionary.
- Create a working copy of the dataset.
- Clean the data in the working file.
- Create an analysis file.

OVERVIEW OF STEPS IN ANALYZING NCD DATA

The focus of this and subsequent data analysis modules is on analyzing NCD data. In order to accurately create an analysis plan, you should be familiar with the steps in analyzing NCD data, for example, from a large complex survey. The steps are as follows:

1. Conduct univariable analyses:

i. Review characteristics of the population

Describe the sample population by person, place, and time characteristics. Summarize variables using population-level frequencies and calculate stratified frequencies across important sub-groups (if any). Determine the frequency distribution of these characteristics.

ii. **Determine frequency of outcome variables**

The objective of many surveys is to estimate the prevalence of one or more health-related behaviors, practices, or outcomes, such as seat-belt use, smoking, hypertension, or diabetes. Univariable analyses should include the frequency distribution of these variables and also confidence intervals for the prevalence estimates.

2. Conduct bivariable analyses:

i. **Display the data in two-variable tables:**

Sometimes the characterization of the population can be extended to two-variable tables, such as age by sex. For surveys in which the objective is to estimate prevalence, the data are often analyzed by population characteristics. For example, you can use two-variable tables to determine whether the prevalence varies by sex or education level. For analytic studies in which the objective is to quantify associations between exposures and outcomes, the two-variable table displays the core result, with rows representing levels of exposure and columns representing presence or absence of the outcome.

ii. **Compute and interpret measures of association:**

Determine the magnitude of association between an exposure variable and an outcome variable. If there are two or more populations, consider comparing their demographic data to determine whether they were different before the study/analysis was conducted.

iii. **Calculate confidence intervals and/or statistical significance:**

Utilize confidence intervals to quantify the variability of the data in your analysis. Use t-tests for continuous data, chi-square tests for categorical data, and other statistical tests as appropriate for the data to determine whether the results are “statistically significant.”

iv. **Assess for effect measure modification:**

Effect measure modification (also known as “effect modification”) is present when an effect measure such as sex, age or geographic location is different at several levels in an exposure-disease relationship. This is evaluated through statistical assessment of interaction between variables.

v. **Assess the effect of potential confounders:**

Confounding is an apparent association between disease and exposure resulting from a third factor that was not considered. A confounder is an independent risk factor for the disease that also happens to be associated with the exposure variable under consideration.

3. Conduct multivariable analyses:

Use your literature review and previous experience to decide on a multivariable analysis and modeling technique to address the hypotheses presented. Utilize the results of the bivariable analysis in implementing your modeling strategy to determine a final model or set of models that best explain your data.

TYPES OF STATISTICAL DATA

Before learning how to create an analysis plan, let us review common types of statistical data by completing the exercise below.



Activity

Practice Exercise: Review of Statistical Data

Instructions:

1. Fill in the blanks below.
2. Check your work by reviewing the answers in Appendix A.

1. Discrete (noncontinuous) data are:

2. Two types of discrete data are:

a. Nominal: (define) _____

i. Nominal data with just two values can be called

ii. Example:

iii. Nominal data can be assigned a

_____ to facilitate analysis.

b. Ordinal:

(define) _____

i. Example: _____

3. Continuous (scale) data are:

4. Two types of continuous data
are: _____

a. Interval (define):

i. Example:

b. Ratio (define):

Example:

Section 2: Analysis Plan

OVERVIEW

An analysis plan is a document you will develop in advance to guide data analysis. The analysis plan usually contains:

- research question(s) and/or hypotheses, if any,
- dataset(s) to be used,
- inclusion/exclusion criteria (e.g., if data only for adults or only for children will be analyzed),
- variables to be used in the main analysis (the main exposure, outcome, and stratifying variables),
- statistical methods and software to be used, and,
- key table shells (univariable, bivariable, and stratified).

Your protocol document should also contain much of this information. (See the **Developing a Protocol** module for more detail.)

Base your analysis plan on the question(s) you need to answer, the information you want to communicate, and the data you have. To do this, you should know where you are starting from (datasets) and where you need to get to (final report). Be specific in deciding what categories to use, for example, age, duration of treatment.

RESEARCH QUESTIONS AND/OR HYPOTHESES

Determine the general research topic (or scope of study) or the questions you need to answer in the analysis. If you are analyzing data in response to a request, determine who needs the answers to the questions, what additional information they need, how often they need the information, and the format in which they need the information.

DATASET(S) TO BE USED

In NCDs, it is common to use large datasets and conduct secondary data analysis. The size of the dataset (or database) depends on the number of records and variables. Commonly used datasets include:

- vital registration (number of deaths, cause of death for a country),
- demographic health surveys (DHS) used in low and middle income countries,
- WHO STEPS survey,

- the National Health and Nutritional Examination survey (NHANES - U.S.), and,
- the Behavioral Risk Factor Surveillance System (BRFSS - U.S., Jordan).

The databases typically are representative of a population either through a census (all persons included) or a sample (number of people selected to represent the population). For example, NHANES 1999–2000 interviewed 9,965 persons in the United States, and the database includes hundreds of variables. Before attempting data analysis for large datasets, it is very important you locate the survey sampling methodology, questionnaire, data variable dictionary and any other supporting documentation.



Tip

Because you most likely did not create the dataset, you must take the time to understand the dataset in its entirety.

Some of the questions you should answer about the dataset are:

- Who owns the database?
- How can you get access to the database?
- Do you need permission to use the database?
- Does the database cost anything to use?
- Are there rules about storing the database?
- What was the purpose of the study?
- What are the study hypotheses?
- What methods were used to identify (select) the population under study and gather information from them?
- How was the data collected, entered, and checked for quality control?
- In what program (e.g., MS Access/Epi Info, Excel, SQL, etc.) and in what format is the dataset stored (e.g., text, ASCII, comma-delimited, etc.)?
- How many records are in the dataset?
- Were weights used?²

² Use weights to account for complex survey design (including oversampling), survey non-response, and post-stratification. When a sample is weighted, it is representative of the population.

- What is the number of observations?

Determine the original purpose of the data and the sponsor or collector of the data. Then determine the study design and methods. Identify whether the data include:

- all persons in the population of interest (census),
- a sample representative of the population (e.g. probability simple random sample, random sample or cluster sampling), or,
- a sample not representative of the population (e.g. non-probability convenience sampling or purposive sampling).

Determine if the dataset contains the variables you need to answer the research questions. Assess how complete and recent are the data.

Determine if you need to conduct a new survey to obtain the required data.

Verify that appropriate instruments were used to collect the data. Keep the questionnaire and codebook (data dictionary) accessible. You can use the data dictionary to learn the coding scheme and the variable names. (In the **Managing Data** module you will learn how to create a data dictionary.)

INCLUSION / EXCLUSION CRITERIA

Describe the criteria you will use to determine which records to analyze. For example, if you have data from an entire country or region but you work in a particular district, your inclusion criteria might include “all records of participants residing in District X.” Similarly, if you are assigned to the Diabetes Unit and you are analyzing hospital discharge data, your inclusion criteria might be “all hospital discharge records with ICD-10 codes E10 to E14.” You might exclude readmissions within 3 days of a previous discharge (which is likely a continuation of the previous problem or a complication from the previous hospitalization rather than a new episode). If your intention is to look at discharge planning, you would exclude any patient that died while hospitalized.

VARIABLES TO BE USED IN THE MAIN ANALYSIS

The Analysis Plan should contain a list of variables³ to analyze that will be kept in the analysis file (a computer file derived from the original data). For

³ *The listing of variables (i.e., data dictionary) is taught in the **Managing Data** module.*

example, if the original file contains information about income, but your analysis does not need to include income, then the analysis file would not include the income variable.

You will also list variables that are not in the original dataset but should be calculated. For example, if the hospital discharge dataset contains date of admission, date of discharge, and date of birth, but it does not include length of stay, then you need to calculate that variable. You would list the name of the variable (“HospDays”), type of variable (integer), its explanation (“number of days in hospital”), and the fact that it is a calculated variable (calculated: HospDisch – HospAdmit +1). The key outcome variables should also be flagged or listed.

STATISTICAL METHODS AND SOFTWARE

There are different statistical methods you will use depending on the research questions. For example, if you want to estimate the prevalence of a behavioral risk factor such as smoking or an outcome such as hypertension, you would first conduct a univariable analysis, then stratify by subgroups. If you need to determine the magnitude of association between an exposure variable and an outcome variable, you will conduct bivariable analysis.

There are many quantitative statistical software packages to use for the analysis. Some examples are:

- SPSS
- STATA
- SAS
- SUDAAN
- Epi Info

Suppose, for example, you are planning to conduct descriptive analysis on the most recent BRFSS study. Because analysis of BRFSS can involve weights for clustering of samples, you will likely need statistical software that can account for this weighting. You will also need a person with training in complex analyses or someone who can help you learn these analysis techniques.



Stop

Let the facilitator or mentor know you are ready for the group discussion.

KEY POINTS TO REMEMBER

Use the space below to record any key points from the facilitator-led discussion:



Activity

Practice Exercise #1 (Estimated Time: 45 minutes)

Hypertension case study

The past few decades have brought a new global phenomenon called the “nutrition transition” in many low and middle income countries⁴. This transition includes a large shift from traditional diets and lifestyles to one

⁴ Popkin, Barry. (2002) *Stages of the Nutrition Transition: Dynamic Global Shifts Appear to be Accelerating*. Available online as of 5/7/2008 at:
http://www.cgdev.org/doc/events/9.10.07/Barry_Popkin_Presentation.pdf

increasingly composed of pre-packaged and processed foods along with sedentary lifestyles. Infectious diseases remain a critical public health priority for parts of the world. (Moore, et al.; WHO, 2002) However, many countries now confront a ‘double disease burden’ with this transition (Yach, et al., 2004) as rates of noncommunicable disease (NCDs), such as diabetes, cardiovascular disease, and cancers account for more than half of the global burden of disease⁵ in both developing and developed countries, (PAHO WHA resolution, 2006). This is, in part, facilitated by global migration into urban settings. In 2005, one-half of the world’s population lived in cities. Most global regions experienced this demographic shift from rural to urban – one of the most rapid in human history.

NCDs are responsible for more than 60% of all deaths worldwide, with more than 80% of NCD-related deaths occurring in low- and middle-income countries (LMICs). (WHO 2011a) Nearly one-third of NCD-related deaths in LMICs occur before age 60. This is compared to only 20% of NCD-related deaths in high-income countries that occur before age 60. NCDs also account for 48% of Disability Adjusted Life Years—DALYs⁶, which pose a challenge for both development and productivity in countries around the world. The burden of NCDs is expected to grow as both the world population and the proportion of persons 60 years and older continue to increase; NCDs disproportionately affect this age group.

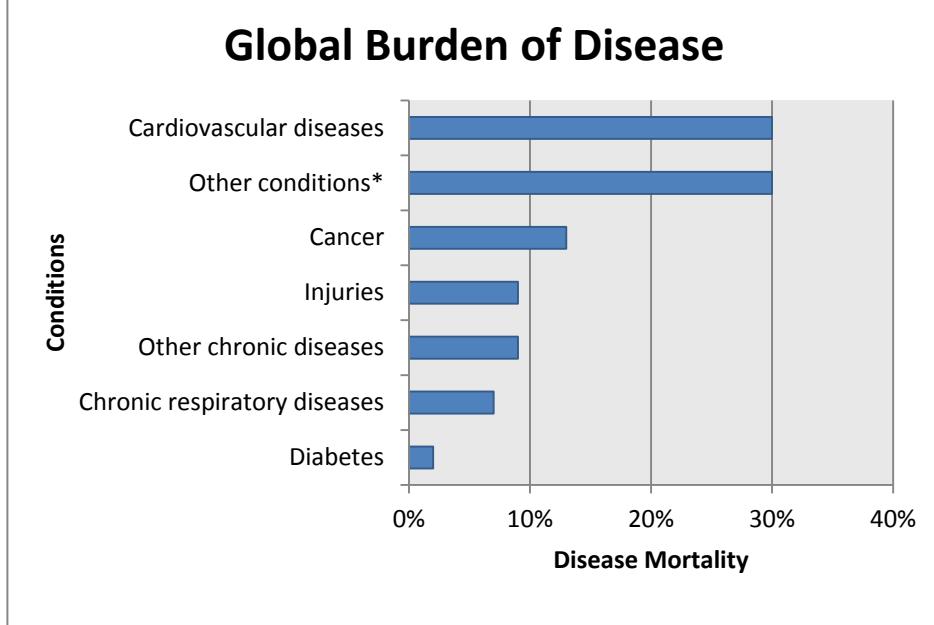
Over the past decade cardiovascular diseases (CVD)⁷ are the single largest cause of mortality worldwide. This represents nearly 30% of all

⁵ Yach D, Hawkes C, Gould C, Hofman KJ. *The Global Burden of Chronic Diseases: Overcoming Impediments to Prevention and Control*. JAMA. 2004;291(21):2616-2622. doi:10.1001/jama.291.21.2616.

⁶ Disability-Adjusted Life Years (DALYS) is the sum of years of potential life lost due to premature mortality and the years of productive life lost due to disability

⁷ **Cardiovascular disease (CVD)** refers to a group of diseases involving the heart, blood vessels, or the sequelae of poor blood supply due to a diseased vascular supply. Over 82% of the mortality burden is caused by ischaemic or coronary heart disease (IHD), stroke (both hemorrhagic and ischaemic), hypertensive heart disease or congestive heart failure (CHF). This varies significantly by global region.

deaths and about 50% of all NCDs (WHO, 2011a). In 2008, CVD caused an estimated 17 million deaths and led to 151 million DALYs. Common behavioral risk factors, including tobacco use, physical inactivity, unhealthy diet and the harmful use of alcohol, are responsible for approximately 80% of the global CVD burden.⁸



Source: Adapted from *Global health risks: mortality and burden of disease attributable to selected major risks*. Geneva, World Health Organization, 2009.

*Includes communicable diseases, maternal and prenatal conditions, and nutritional deficiencies.

Raised blood pressure, or hypertension⁹, is the leading risk factor for mortality and is ranked third as a cause of disability-adjusted life-years.¹⁰ It

⁸ Gaziano TA, Bitton A, Anand S, Abrahams-Gessel S, Murphy A. *Curr Probl Cardiol*. 2010 Feb;35(2):72-115.

⁹ Hypertension is defined as blood pressure 140/90mmHg or above most of the time. If blood pressure is > 120/80 mmHg and < 140/90 mmHg, it is called pre-hypertension. Normal blood pressure is 120/80. (Chobanian, A, Bakris, G. et al. *The Seventh Report of*

has been estimated that hypertension resulted in 51% of stroke deaths and 45% of coronary heart disease deaths in 2008.¹¹ Mean blood pressure has decreased significantly in nearly all high-income countries due to widespread diagnosis and treatment along with access to low-cost medications. In contrast, mean blood pressure has been stable or increasing in most African countries; approximately 40% (and up to 50%) of adults in many of these countries are estimated to have high blood pressure. Most of these people remain undiagnosed, although many could be treated with low-cost medications; this would significantly reduce the risk of death and disability from heart disease and stroke.¹²

Effective prevention strategies for NCDs, and specifically for reducing the burden of hypertension and CVDs, do exist. For example, efforts to reduce sodium consumption have been identified as a cost-effective means to reducing and reversing hypertension. Increasing awareness and education of the consumer through campaigns to encourage dietary change within households with low-sodium alternatives as well as use of salt substitutes is one approach. In addition, working with industry to encourage voluntary reduction of salt content of processed foods and condiments by manufacturers is another proven strategy. Recent estimates indicate that implementation of a salt reduction program could avert 8.5 million deaths globally. Combining this with implementing the WHO Framework Convention on Tobacco Control could save an additional 5.5 million lives. This would be at a cost of less than USD \$0.40 per person per year in low-income and lower middle-income countries, and

the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure. JAMA. 2003;289:2560-91.

www.nhlbi.nih.gov/guidelines/hypertension/jnc7full.pdf)

¹⁰ Asaria P, Chisholm D, Mathers C, Ezzati M, Beaglehole R. Chronic disease prevention: health effects and financial costs of strategies to reduce salt intake and control tobacco use. *The Lancet* 2007; 370(9604):2044-2053.

¹¹ *World Health Statistics: A Snapshot of Global Health.* World Health Organization, 2012. Geneva, Switzerland.

http://who.int/gho/publications/world_health_statistics/2012/en/index.html

¹² Yach D, Hawkes C, Gould C, Hofman KJ. *The Global Burden of Chronic Diseases: Overcoming Impediments to Prevention and Control.* JAMA. 2004;291(21):2616-2622. doi:10.1001/jama.291.21.2616.

USD \$0.50 – 1.00 per person per year in upper middle-income countries (as of 2005).¹³

Effective strategies, however, require specific data on risk factors in order to set priorities and develop and monitor interventions.

How One Country is Addressing the Burden of CVD

Country X is a rapidly modernizing nation of 10 million people, with a growing middle-class. Vital statistics indicate a gradually increasing overall life expectancy but a surprisingly high rate of deaths due to cardiovascular disease, specifically stroke. Health officers from two southern provinces report a steady increase in the use of renal dialysis services. This is placing serious constraints on the regional health service budgets. The traditional diet of the country is generally rich in fresh vegetables, whole grains, and healthy oils; however, the younger and middle-aged segments of the population have rapidly increased their consumption of meals out of the home, which are typically higher in fat, salt, and processed foods. There has been a tremendous expansion of restaurants that serve meals not typical of the traditional diet in the country.

The National Health Service is able to provide limited specialty care services; however, it is able to provide affordable primary care and basic medications to citizens based on family income.

A recently assembled panel of health care leaders in the country gave recommendations to the newly appointed Minister of Health (MoH). Their report highlighted concerns about cardiovascular disease, stroke, chronic renal disease, and other NCD risk factors such as physical inactivity and changing diet. However, the report cited particular concerns about the lack of basic information on **hypertension**, a potential factor underlying these other conditions.

You are the leader of the MoH's Chronic Disease Surveillance Unit. The

¹³ Yach D, Hawkes C, Gould C, Hofman KJ. *The Global Burden of Chronic Diseases: Overcoming Impediments to Prevention and Control*. JAMA. 2004;291(21):2616-2622. doi:10.1001/jama.291.21.2616.

MoH has asked you to analyze national health survey data on **hypertension**, which has been collected every two years over the past decade, and other provincial level hospital data and report on the findings. The most recent data were collected last year.

The MoH wants to provide the report to the national and provincial decision-makers so they can better understand the magnitude of the burden of disease of hypertension and the key determinants and underlying factors that are affecting this public health burden. With this information, the MoH is hoping to target resources and support evidence-based actions and policies to improve the health of the population.

Fill in the following sections based on the information from the case study and the questionnaire:

Research question(s) and/or hypotheses:

Dataset(s) to be used:

Inclusion/exclusion criteria:

Variables to be used in the main analysis: (List 3 or 4 outcomes and exposure variables. You will create a data dictionary in another

module.)

Statistical methods and software to be used: (Answer according to the software you use in your country.)

OVERVIEW OF PREPARING TABLE SHELLS

Good epidemiologic practice dictates that you plan data analysis methods before you conduct analyses. When analyzing large datasets, you will conduct three types of analyses: univariable, bivariable, and multivariable data analyses. These analyses will logically proceed from simple

(univariable) analyses to complex (multivariable). This allows you to understand your data and to make decisions about the next steps for each type of analysis (i.e., you will use the results of univariable and bivariable analyses to select methods and variables to evaluate in the multivariable analysis(es)). You will prepare tables in advance to help you think through analyses, organize, and present the descriptive information.

You will use the table shells created in the analysis plan to analyze the data; however, remain flexible if you decide to group categories differently or pursue interesting, unanticipated findings that help address the hypotheses being tested. Also, analysis often uncovers additional errors. Be sure to document and correct these errors for the dataset as a whole to help ease the burden on future users of this data.

TYPES OF TABLE SHELLS

Your analysis plan will consist of table shells¹⁴ that should be ready for publication except for the data. Each table shell will contain a title, category labels, analytic techniques to be used, expected format for results, but **no data**. You will create table shells to help you prepare for the following analysis activities:

- conducting descriptive analysis,
- calculating measures of association (OR),
- calculating confidence intervals,
- conducting statistical testing,
- assessing potential effect modification and confounding, and,
- conducting multivariable analysis.

Note: Because this and subsequent modules for this course focus on basic analysis, you will not learn about preparing tables for multivariable analyses.

Univariable analysis

A univariable data analysis is when you analyze one variable at a time in a dataset; this is sometimes referred to as descriptive analysis. For this type of analysis, you examine the range, mean, median, and mode of each continuous variable or the range and frequency distribution of discrete variables. Depending on the questions you need answered, this analysis

¹⁴ See Appendix B for an example of study objectives, a hypothesis and table shells.

can provide you with information about the factors of person, place, and time in the population of interest such as:

- the characteristics of the population, such as age, gender, where they live (e.g., urban or rural),
- the prevalence of risk factors among the population,
- the prevalence of the disease, outcomes, or exposures in the population, or,
- when the events of interest occurred, such as monthly or yearly.

When you prepare for descriptive analysis you are likely to create new variables, recode variables, and possibly combine variables. For example, if you have collected date of birth information, you may want to create one variable reflecting age (continuous), or another variable reflecting age category (for example, 10 – 19). Or, it may be part of your study objectives to create a variable reflecting body mass index calculated from two separate variables representing height and weight. Also consider measures of association you may want to calculate later, and recode exposure or outcome variables, as appropriate.

For example, if a study objective was to determine current tobacco use among adults by gender, educational level, and location, an example of a table shell you may create is:

Education Level of Adults

Highest Educational Achievement	Number	Percent
None		
Primary school		
Secondary school		
Post-secondary school		
Total		



Activity

Use the space below to create another example of a univariable table shell for the study objective on the previous page. Review with a colleague or facilitator.

Bivariable analyses

Bivariable data analysis involves analyzing the relationship between two variables. You can conduct bivariable analysis to test simple hypotheses of association and causality. Some examples of this technique include comparing the outcome of interest in terms of:

- demographic characteristics (e.g., comparing differences in age, gender, ethnicity, income, or location between cases and controls), and,
- exposure characteristics (e.g., comparing differences in drug use, environmental exposure, diet, exposure to other ill persons, family history of disease, animal/insect exposure between cases and controls).

An example of a bivariable table shell:

Tobacco Use by Education Level

Exposure Variable: <i>Highest Educational Achievement</i>	Outcome Variable: <i>Tobacco Use</i>		Total
	Yes	No	
None	_____ (%)	_____ (%)	
Primary school	_____ (%)	_____ (%)	
Secondary school	_____ (%)	_____ (%)	
Post-secondary school	_____ (%)	_____ (%)	
Total			



Activity

Use the space below to create another example of a bivariable table shell for the tobacco use study. Review with a colleague or facilitator.

Measures of Association, Confidence Intervals, and Statistical Testing

Epidemiologic studies often explore the association between exposures or risk factors and the disease or outcome of interest in a study population. Many initial studies are case-control and cross-sectional studies that provide information on prevalence of exposures and prevalence of outcomes at a given moment in time. With these studies, you can estimate the probability that a person who has the outcome also has exposure to a risk factor. This can be done through prevalence ratio (PR) for a cross sectional study or prevalence odds ratio (POR) for a case-control study. With chronic diseases, you may have a dataset that is collected as a cohort over a longer period of time that you can analyze using the risk ratios or rate ratios (RR).

After describing characteristics, exposures, and outcomes in a study population and its sub-groups, the next step is to consider whether characteristics differ by a statistically significant margin between groups. The bivariable and multivariable analysis methods used to assess these characteristics depend on the type of variable and the research question or hypothesis.

Use statistical tests to assess associations between variables in which you are interested in finding a statistically significant association. Statistical tests include t-tests for continuous data, chi-square (χ^2) tests for categorical data, ANOVA for assessing a continuous variable within categories, and a correlation coefficient to assess correlation between two continuous variables. (**Note:** For this training, we will only review t-tests and chi-square tests.)

The following is an example of a table shell for preparing to calculate measures of association and conducting statistical testing and confidence intervals:

Exposure Variable: <i>Highest Educational Achievement</i>	Outcome Variable: <u>Tobacco use</u>				
	Yes	No	Total	PR (95% CI)	POR (95% CI)
None	_____ (%)	_____ (%)		1.0 _____ —	1.0 _____
Primary school	_____ (%)	_____ (%)		_____ —	_____
Secondary school	_____ (%)	_____ (%)		_____	_____
Post-secondary school (referent)	_____ (%)	_____ (%)		_____ —	_____
Total					
X² = _____, df = _____, p = _____,					

Stratified Analysis to Assess for Effect Measure Modification and Confounding

After you assess the strength of a two-variable exposure-outcome relationship, you will next stratify by other variables. The measure of association can be impacted by other variables in the dataset, often called covariates or third variables. These variables may be modifiers of the measure of association or confounders.

Effect measure modification (also known as “effect modification” occurs when the measure of association (e.g., the odds ratio, prevalence ratio) between an exposure and an outcome is different depending on the value of a third variable. This means the effect of the exposure on the outcome is different for different levels of the third variable. When you stratify the third variable you can detect effect measure modification. Effect modification is present when the stratum-specific measures of association different from each other. For example, an exposure may substantially increase a woman’s risk of a particular disease but have little or no such effect in men.

Recall that confounders are variables that distort the relationship between the exposure and the outcome. Confounding threatens the validity of an epidemiologic study since it can lead to false conclusions regarding the true relationship between an exposure and outcome. Confounding can either overestimate or underestimate the true magnitude of the measure of association between an exposure and outcome.

The following is an example of a table shell to prepare for stratified analysis (i.e., confounding and effect modification):

Exposure Variable:	Outcome Variable: <u>Tobacco use</u>				
<u>Highest Educational achievement</u>	Yes	No	Total	PR (95% CI)	POR (95% CI)
None	____ (%)	____ (%)		1.0 _____	1.0 _____
Male	____ (%)	____ (%)		_____	_____
Female (referent) ¹⁵	____ (%)	____ (%)		_____	_____
Primary school	____ (%)	____ (%)		1.0 _____	1.0 _____
Male	____ (%)	____ (%)		_____	_____
Female (referent)	____ (%)	____ (%)		_____	_____
Secondary school	____ (%)	____ (%)		1.0 _____	1.0 _____
Male	____ (%)	____ (%)		_____	_____
Female (referent)	____ (%)	____ (%)		_____	_____
Post-secondary school	____ (%)	____ (%)		1.0 _____	1.0 _____
Male	____ (%)	____ (%)		_____	_____
Female (referent)	____ (%)	____ (%)		_____	_____
Total	____ (%)	____ (%)			
X² = _____, df = _____, p = _____,					

¹⁵ “Female” was selected as the referent category because they are less likely to smoke than males.



Stop

Let the facilitator or mentor know you are ready for the group discussion.

KEY POINTS TO REMEMBER

Use the space below to record any key points from the facilitator-led discussion:



Activity

Practice Exercise #1 (Estimated Time: 45 minutes)

Background:

For this exercise, you will work individually, in pairs, or in a small group to create table shells.

Instructions:

1. Read the case study information in the three sections below and answer the questions that follow.

Ask a facilitator to review your work.

Section 1 - Case Study:

As you are familiarizing yourself with the national health survey dataset, you receive an inquiry from the Minister of Health (MoH). The MoH would like additional information about the national health survey and estimates of the current adult population with hypertension.

1. Based on the case study information, what type of univariable analyses will you conduct? Use the space on the following page to create at least three table shells to prepare for univariable analyses.
2. Based on the case study information, what type of bivariable analyses will you conduct? Use the space below to create at least two table shells to prepare for bivariable analyses:

Section 2 - Case Study:

Using the national data available as requested by the MoH, you are interested in better understanding the magnitude and risk factors related to cardiovascular diseases, and would like to explore and test these associations.

1. Use the following template to list the variable pairs for which you will test a statistical association.

Statistical Test	Variables to Assess
Chi-square	_____ vs. _____
t-test	_____ by _____

2. Fill out the following tables to prepare for calculating measures of

association.

Exposure Variable:	Outcome Variable: _____	
_____	Yes	No

PR =

POR =

χ^2 =

Exposure Variable:	Outcome Variable: _____	
_____	Yes	No

PR =

POR =

χ^2 =

Exposure Variable:	Outcome Variable: _____	
_____	Yes	No

PR =

POR =

χ^2 =

Section 3 - Case Study:

To assess for confounding, you need to consider a research question. Obesity estimates have been rising in your country. You consider this to be a potential contributing factor to the reported rising estimates of hypertension. Consider: Is obesity a risk factor for hypertension? What are potential confounders in the relationship between obesity and hypertension?

(Note: When assessing confounding and effect modification, consider stratification of variables (e.g., age group, gender, etc.) to assess the primary relationship (i.e., obesity and hypertension). Stratification allows you to observe relationships beyond the crude association.)

1. List at least one potential confounder for the obesity-hypertension analysis.

What table shells do you need to prepare to assess for confounding and effect modification? Use the space below to create at least two table shells.



Stop

Take out the activity workbook. Let your facilitator or mentor know you
ready to begin the skill assessment.

Resources

For more information on topics found within this workbook:

Rothman K, Greenland S, Lash T. Modern Epidemiology. 3rd Ed. Philadelphia, PA: Lippincott Williams & Wilkins; 2008.

Jekel J, Katz D, Wild D, Elmore K. Epidemiology, Biostatistics and Preventive Medicine. 3rd Ed. Philadelphia, PA: Saunders; 2007.

Dicker RC. Analyzing and Interpreting Data. In: Gregg MB, ed. Field Epidemiology, 3rd ed. New York: Oxford U. Press; 2008.

For information about using PR and POR:

Thompson ML, Myers JE, Kriebel D. Prevalence odds ratio or prevalence ratio in the analysis of cross sectional data: what is to be done? Occup Environ Med. 1998 Apr;55(4):272-7.

<http://www.ncbi.nlm.nih.gov/pubmed/9624282>

Pearce N. Effect measures in prevalence studies. Environ Health Perspect. 2004 Jul;112(10):1047-50.

<http://www.ncbi.nlm.nih.gov/pubmed/15238274?dopt=abstract>

Reichenheim ME, Coutinho ES. Measures and models for causal inference in cross-sectional studies: arguments for the appropriateness of the prevalence odds ratio and related logistic regression. BMC Med Res Methodol. 2010 Jul 15;10:66.

<http://www.ncbi.nlm.nih.gov/pubmed/20633293>

Santos CA, Fiaccone RL, Oliveira NF, Cunha S, Barreto ML, do Carmo MB, Moncayo AL, Rodrigues LC, Cooper PJ, Amorim LD. Estimating adjusted prevalence ratio in clustered cross-sectional epidemiological data. BMC Med Res Methodol. 2008 Dec 16;8:80.

<http://www.ncbi.nlm.nih.gov/pubmed/19087281>

Appendices

Appendix A

Answers to Quiz:

1. **Discrete (Noncontinuous)**: data values that fit into distinct categories.
2. Two types of discrete data are nominal and ordinal
 - b. **Nominal**: data values are nonnumeric group labels with mutually exclusive categories; they cannot be ranked.
 - i. Nominal data with just two values can be called **dichotomous**.
 - ii. Example: males/females or married-Yes/Married-No.
 - iii. Nominal data can be assigned a code in the form of a number to facilitate analysis.
 - c. **Ordinal**: data values are categorical and can be ranked (i.e., put in order) or have a rating scale.
 - i. Example: strongly disagree to strongly agree may be assigned values from 1 to 5.
3. **Continuous (scale)**: numeric data values that take on any value within a range. For example, age, height, and weight are continuous data.
4. Two types of continuous data are: interval and ratio
 - a. **Interval**: data values are ranged in a real interval and the data has an arbitrary zero. The difference between two values are meaningful, however, the ratio of two interval data is not meaningful.
 - i. Example: temperature, dates.
 - b. **Ratio**: a comparison of numeric data values. Both the difference between the values and ratio of the values are meaningful.
 - i. Example: height, weight, and age.

Appendix B

The following study objectives, hypotheses and table shells have been adapted from the Guatemala FETP: **Study on prevalence and factors associated to tobacco consumption among adolescent students from middle level schools in Antigua Guatemala, Sacatepéquez, Guatemala, 2012.**

1. General objective

Estimate the prevalence of tobacco use and its associated factors among students of the middle level that includes the first, second, and third basic grades, and the students are presently attending a private or public educational center during the morning or afternoon educational schedules in Antigua Guatemala, Sacatepéquez.

2. Specific objectives

- Estimate the life prevalence and the current prevalence of tobacco consumption.
- Determine the intention to smoke, the access to tobacco, and the exposure to passive smoking among these students.
- Identify other risk factors associated with the use of tobacco.

Null Hypothesis (Ho): That the current prevalence of tobacco consumption among students of the middle level in Antigua Guatemala is equal or lower than the prevalence of tobacco consumption among students of the middle level at the national level.

Table No. 17

Tobacco consumption among students in Antigua Guatemala,
Municipality of Sacatepéquez, Guatemala, 2012

Tobacco Consumption

	Number of Cases	Percentage
Yes		
No		
No response		
Total		
> 18		
TOTAL		

Table No. 39

Tobacco consumption by gender among
students in Antigua Guatemala,
Municipality of Sacatepéquez, Guatemala, 2012

Tobacco consumption

Gender	Number of cases	Percentage
Male		
Female		
Total		

Variable of Exposure: <u>Friends that smoke</u>	Variable of Result: <u>Intention to smoke</u>	
	Yes	No
Yes		
No		

Variable of Exposure: <u>Others smoking at home</u>	Variable of Result: <u>Smoker</u>	
Yes	Yes	No
No		