RRFSS Analysis
Orientation Manual

RRFSS
Rapid Risk Factor Surveillance System

2016 EDITION

Developed and Edited by
RRFSS Representatives and RRFSS Coordinator
November, 2015

Table of Contents

Welcome!

Welcome to the Rapid Risk Factor Surveillance System (RRFSS). As a new representative it can be confusing and a little overwhelming to become familiar with the many facets of RRFSS. Knowing this, it seemed that something was needed to ease the transition for new RRFSS representatives and help get them up and running more quickly. This manual is intended to address just those needs. While there are many very informative RRFSS documents that already exist that will be referred to in this manual, this publication differs somewhat in that it seeks to pull together a number of those pieces and highlight the parts most useful to the new representative. We would encourage you to read this manual thoroughly as an overall orientation and move on to the more technical documents when you have become familiar with the big picture or as needed.

As you read this document, it would be helpful to have the most current versions of the documents shown in Table 1. If you do not already have these, don't hesitate to contact those individuals listed below for this information. As well, make sure you have access to the RRFSS website also indicated below.

You may also want to keep the glossary found at the end of the manual handy. Not being familiar with terminology, jargon, and acronyms can make any new process alienating and often those who've been at it for a while forget that they're taking a lot of knowledge for granted. This glossary should assist you with beginning to speak "RRFSS"!

Table 1: Important RRFSS Resources

RESOURCE	DESCRIPTION	WHERE DO I FIND IT?
RRFSS Manual of Operations (MOO)	Guidelines and procedures for RRFSS	Lynne Russell RRFSS Coordinator 416-736-2100 ext 22556 lynnerussell@rrfss.ca
RRFSS Contact List	Participating RRFSS members and personnel, and relevant ISR personnel	
Annual Technical Documentation	An annual report issued by ISR (usually in July) summarizing the sample design, data collection, and data processing, as well as the RRFSS questionnaire	Liza Mercier Institute for Social Research York University 416-736-2100 ext. 20356 lmercier@yorku.ca
RRFSS Website	Open to the general public, the RRFSS website provides general information about RRFSS, as well as local prevalence data,	www.rrfss.ca

	questionnaires, data dictionaries, etc.	
Questionnaire Maps (Q-Maps)	An Excel sheet that lists all RRFSS modules being asked by RRFSS participating health units for that particular wave.	Liza sends the maps a week or so before the start of the new wave. RRFSS reps need to check them to make sure that their module selection for the next wave is correct.
Module Inventory	A current list of all the RRFSS modules and the number of questions per module	Lynne Russell RRFSS Coordinator 416-736-2100 ext 22556 lynnerussell@rrfss.ca
Interview Length	An Excel that Liza sends out that lists the average interview length by health unit for the previous month and for the previous years. Reps need to check that so that they adjust their modules if they are either way under or way over their survey length.	Liza send this out to reps.
Members List	An Excel sheet that Lynne periodically sends out – the sheet has the contact names, phone numbers, fax numbers, and e-mail addresses of all the RRFSS reps and alternates (by health unit) On the e-mail that Lynne sends out with this list, she usually states if there's any changes to the membership (i.e new reps, etc.)	Lynne Russell RRFSS Coordinator 416-736-2100 ext 22556 lynnerussell@rrfss.ca
Working Groups List	Lynne periodically sends out a list of all the working groups and adhoc groups in RRFSS with the associated lead and members	Lynne Russell RRFSS Coordinator 416-736-2100 ext 22556 lynnerussell@rrfss.ca

Brief Overview of RRFSS

The Rapid Risk Factor Surveillance System (RRFSS) is an on-going telephone survey occurring in various public health units across Ontario. If you are a RRFSS representative, then your health unit is part of RRFSS. RRFSS is a random sample of adults aged 18 years and older are interviewed in the area of each participating health unit. Questions tap into risk factors, knowledge, attitudes, and awareness of topics of importance to public health. Data is provided to participating health units three times per year, allowing for highly up to date health status information. The survey is conducted by the Institute for Social Research (ISR) at York University, on behalf of all RRFSS-participating health units.

RRFSS Who's Who?

There are a number of important people who make RRFSS operate smoothly. While your RRFSS Contact List will provide names and contact information for all RRFSS members, staff, and other important persons, it's probably smart to know the following individuals:

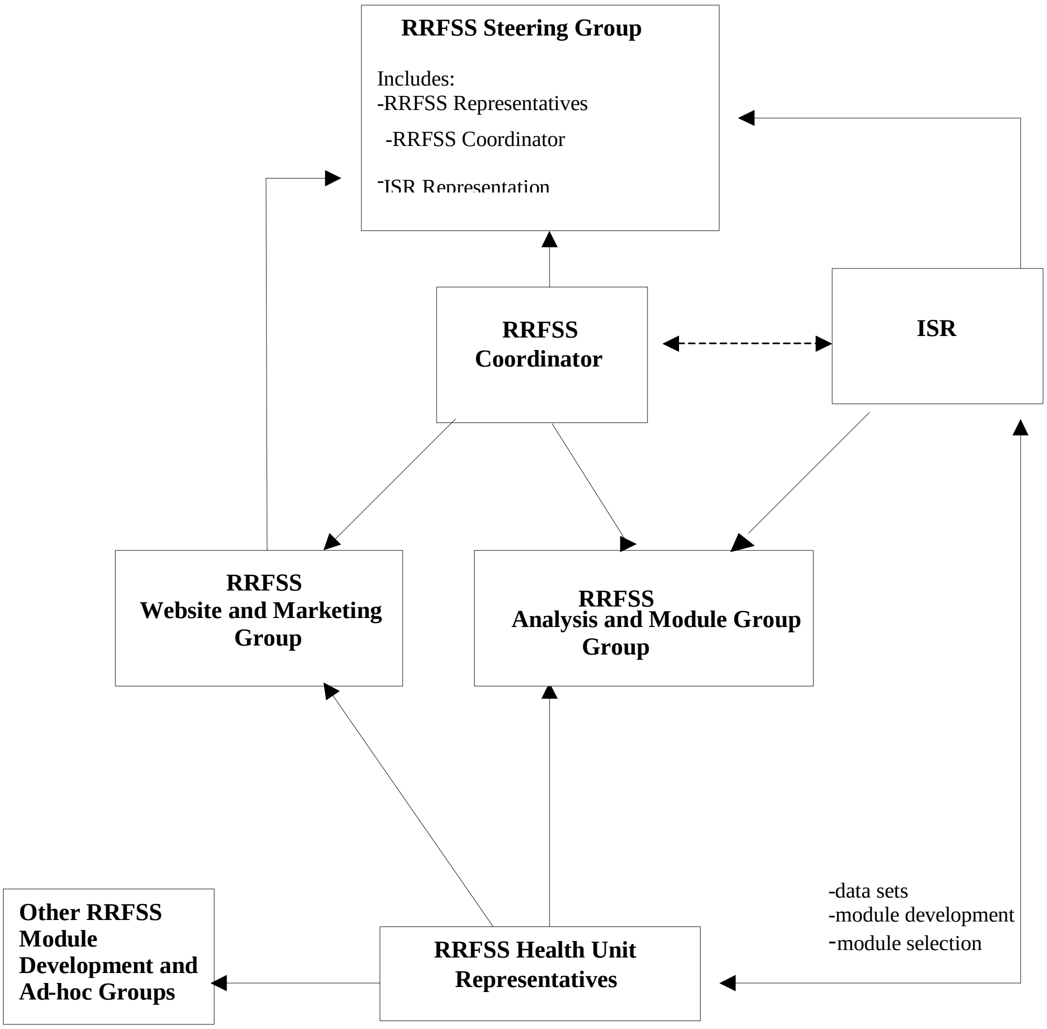
Lynne Russell RRFSS Coordinator 416-736-2100 ext 22556 lynnerussell@rrfss.ca	Lynne is the person with all the administrative knowledge and history of RRFSS. She's most helpful with the following sorts of things: • Current RRFSS Documentation (e.g., MOO, contact list) • RRFSS module info • Minutes for all RRFSS groups Lynne is also a good person for general questions as she can often get you an answer if she doesn't have it already. However, Lynne is not the person to contact for data analysis questions.
Liza Mercier Institute for Social Research York University 416-736-2100 ext. 20356 lmercier@yorku.ca	Liza is the project manager for RRFSS. She most often communicates with RRFSS representatives via e-mail about the following: • Q-maps for surveys • Time estimates • Contract info You can also contact Liza for information about survey protocols, RRFSS contracts, and any phone calls you may receive from potential respondents from your health unit.
Chris Clubine-Ito Institute for Social Research York University 416-736-2100 ext. 77171 cclubine@yorku.ca	Chris will probably be best known to you as the person who emails you the data file. His emails include careful directions on how to obtain data files from the ISR server and if there are any issues with the dataset. Make sure you read the whole e-mail!!! If you were unable to download RRFSS files when they were first

	released, you may contact Chris to gain access to these files.

RRFSS Structure

In terms of organizational structure, RRFSS consist of a number of core groups, all of which have specific roles. While all RRFSS groups are important and take part in decision-making, it is helpful to think of RRFSS in terms of the chart shown below.

RRFSS Organizational Chart



While you can also refer to **SECTION 1.0 TERMS OF REFERENCE** in the **MOO** for more information on the roles and responsibilities of the various groups, the following gives you a brief description of what each group does.

Steering Group

The core responsibility of the Steering Group is to oversee the strategic direction of RRFSS, as determined by the RRFSS membership. In more practical terms, the

Steering Group also provides direction to the RRFSS Coordinator, while acting as the link between ISR.

Analysis and Module Group

If data analysis issues are of concern, it is the Analysis and Module Group that can help. Their role is to advise the Steering Group on analysis issues, making necessary suggestions or recommendations. They also review and respond to analysis concerns that come to them via the Steering Group or from RRFSS reps directly. You should also contact the Analysis and Module Group if you notice any problems with the survey that may impact data analysis (e.g., an incorrect skip pattern).

Website and Marketing Group

This group works to guide and oversee the maintenance and development of the RRFSS Web Site including its structure, function, content, domain name and the resolution of specific website issues. Its role is also to provide guidance on the promotion of RRFSS and the development and implementation of a Marketing and Communication plan.

Other Ad Hoc Groups

At times it is necessary to strike other RRFSS Working Groups on an ad hoc basis.

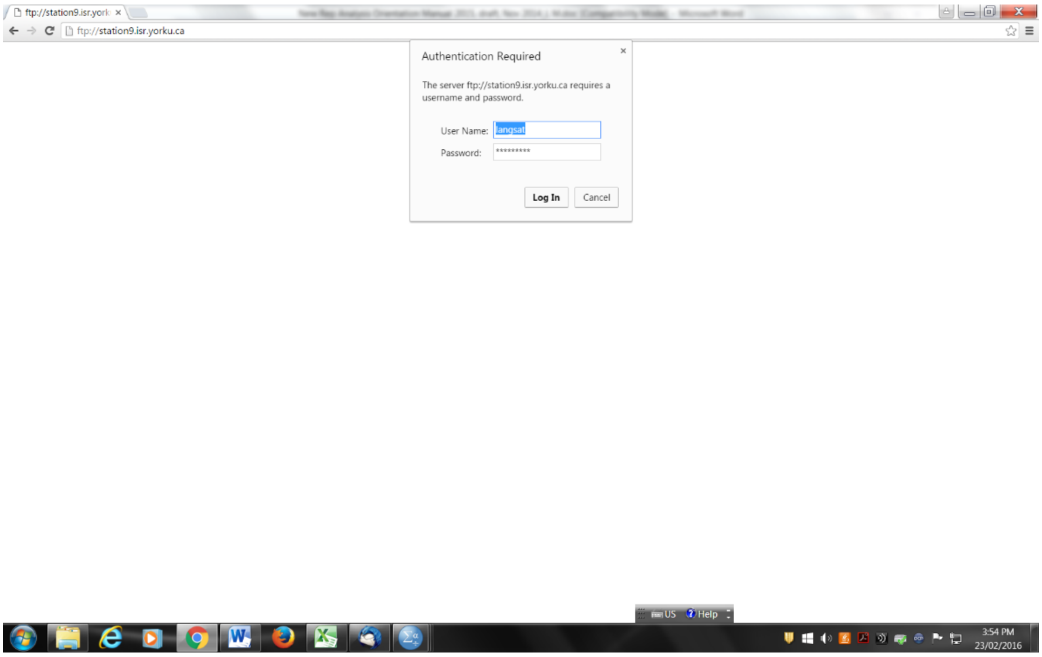
Accessing Data Sets

As a new rep you may be eager to start in on making the most of RRFSS, which means providing up-to-date health status information to your health unit. However, in order to do so, you need to become familiar with your basic tool – your data set. For a quick overview, read on. For more details you can also refer to **2.15 Procedure for Data Sets Access and Security** in the **MOO**.

How do I get data?

RRFSS data is collected on a cyclical basis and it usually takes about 8 weeks from the end of data collection for you to be notified about a completed data set (e.g., data collection for the January to April cycle means notification emails in early July). Emails regarding data are sent from Chris Clubine-Ito (cclubine@yorku.ca) of ISR.

The first email you receive will provide the ftp link to use to begin the data retrieval process (i.e., ftp://langsat@station9.isr.yorku.ca). Your first step is to click on that link or cut and paste it into your browser. Once you have done this, you'll be asked for a username and password which are also included in the email from ISR.



Authentication Required

The server ftp://station9.isr.yorku.ca requires a username and password.

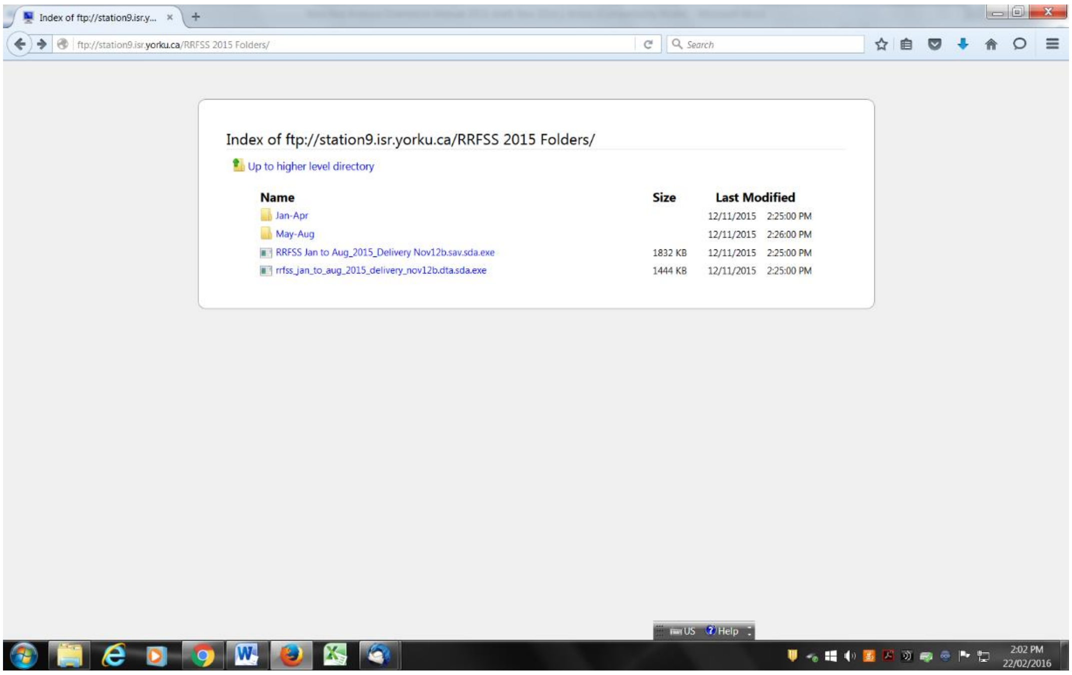
User Name: langsat

Password: *********

Once you have entered the username and password, your browser window will display the current files available for download. Although you may not know what all the files are for, it's a good idea to download all of them. Please note that the password expires one month after the e-mail is sent out, so make sure you download the files as soon as you can, especially the text files

What are all these files?

Now that you have your files you may be wondering what is in them. The figure below shows you the files in a typical download, keeping in mind your icons may look different depending on your operating system and personal preferences. Those shown are the files from the January-August 2015 data collection (Cycles 19 & 20).

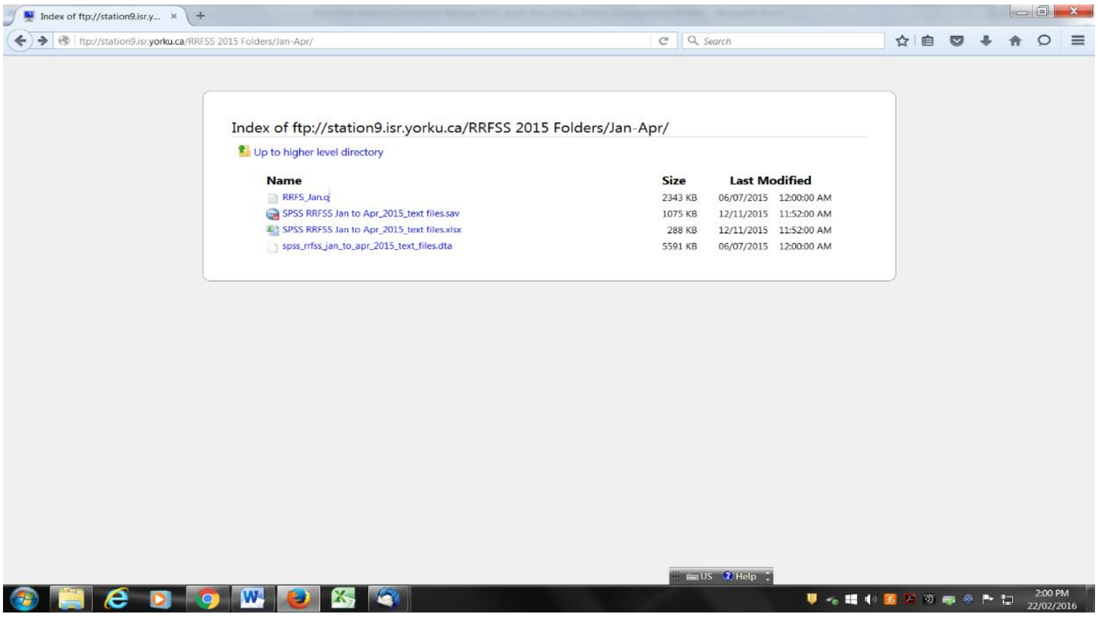


Index of ftp://station9.isr.y...
ftp://station9.isr.yorku.ca/RRFSS 2015 Folders/
Search
2:02 PM
22/02/2016

Index of ftp://station9.isr.yorku.ca/RRFSS 2015 Folders/

Up to higher level directory

Name	Size	Last Modified
Jan-Apr		12/11/2015 2:25:00 PM
May-Aug		12/11/2015 2:26:00 PM
RRFSS Jan to Aug_2015_Delivery Nov12b.sav.sda.exe	1832 KB	12/11/2015 2:25:00 PM
rrfss_jan_to_aug_2015_delivery_nov12b.dta.sda.exe	1444 KB	12/11/2015 2:25:00 PM



Index of ftp://station9.isr.y...
ftp://station9.isr.yorku.ca/RRFSS 2015 Folders/Jan-Apr/
Search
2:00 PM
22/02/2016

Index of ftp://station9.isr.yorku.ca/RRFSS 2015 Folders/Jan-Apr/

Up to higher level directory

Name	Size	Last Modified
RRFS_Jan.q	2343 KB	06/07/2015 12:00:00 AM
SPSS RRFSS Jan to Apr_2015_text files.sav	1075 KB	12/11/2015 11:52:00 AM
SPSS RRFSS Jan to Apr_2015_text files.xlsx	288 KB	12/11/2015 11:52:00 AM
spss_rrfss_jan_to_apr_2015_text_files.dta	5591 KB	06/07/2015 12:00:00 AM

Note: screen shot may not appear exactly as shown above

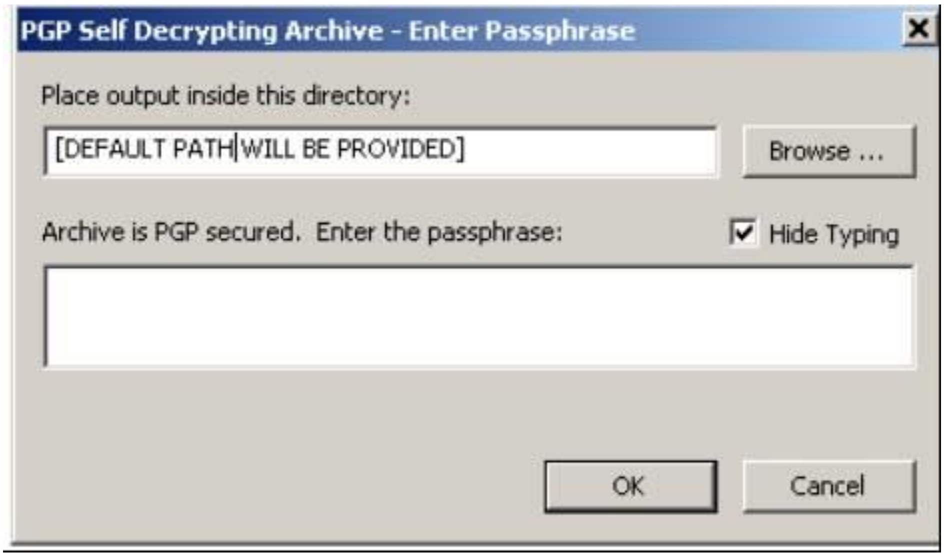
What is in the files is simple if you understand a few things about the naming conventions used, such as:

- Files that end with '_Q' are the CATI questionnaires used for that cycle. They can be opened with Word

- Text files (open-ended answers) are delivered in SPSS (files ending with '_text files.sav'), STATA (files ending with '_text files.dta'), and Excel ((files ending with '_text files.xlsx')

- Data files are delivered in in SPSS (files that end with 'sav.sda.exe') and STATA (files that end with dta.sda.exe)

One last step…
There is one final step before you can use the data. Your data file is sent to you as an encrypted (locked) file. This is to further ensure that only the designated people can access the data. When you try and double click on this file you will be prompted with a screen like that shown below.



PGP Self Decrypting Archive - Enter Passphrase

Place output inside this directory:

[DEFAULT PATH WILL BE PROVIDED]

Archive is PGP secured. Enter the passphrase:

☑ Hide Typing

All you do here is type in the password that was provided in the second email from ISR to unlock the file and you're up and running. If you're unsure what this password is, a quick email to Chris Clubine-Ito (cclubine@yorku.ca) or another RRFSS rep should provide you with the answer. Once unlocked, your data set will open in SPSS (or STATA).

Note: some RRFSS reps prefer to have their data files delivered on a CD or posted to their Health Unit's FTP site.

Organization of the Data Set

Now that you have your data set open, you're confronted with a very large data set with many variables. Understanding the lay of the land, however, should demystify things a bit more. One thing needs to be cleared up right away. As previously stated, RRFSS data is sent out every four months, starting anew each year in January. However, the data that is sent out at the end of each data collection cycle is also CUMULATIVE. In other words, each new data set over a year contains the new data along with all previous data from the same year. Thus, for any given year:

- Data set 1 – contains January-April data
- Data set 2 – contains January-April data, and May-August data
- Data set 3 – contains January-April data, May-August data, and September to December data

By the end of the year, when you receive your 3rd data set, you will have all the RRFSS data for the all participating health units for the preceding year. One good thing about this is that if you miss downloading data at any point mid-year, you need only wait for the next data set to arrive and you have a complete data set.

In addition to understanding which data is contained in the data sets, there are three other useful things to know about in your data set: cycles, health unit ID, and weights.

Cycles

While data is collected and distributed every 4 months, it is more accurate and useful to get in the habit of understanding your data set in terms of cycles. In point of fact, each cycle of data collection can be identified in the data set. For example:

- January-April 2015 data = cycle19
- May-August 2015 data = cycle 20
- September-December 2015 data = cycle 21
- All 2015 data = cycles 19-21

Cycles are numbered sequentially with each new year, beginning where the previous year left off. The variable name for wave in the data set is simply "CYCLE". If you were writing a report you would likely indicate the months of data you are analyzing (e.g., January-April 2015) as it would make more sense to the reader. However, for ease of data analysis you would select cycle 19, which corresponds to the January-April 2015 data.

Cycles are important to know because you may want to select data for only certain months or seasons. For example, say you wanted to analyze the Immunization Flu module for 2015. This is a seasonal module because it is asked only from January to April. This means you only want to select the data from Cycle 19. For datasets prior to 2009, you would need to select the Wave 73 (January, 2007) through Wave 76 (April, 2007). For a list of Cycles/Waves and their corresponding months/years, go to the RRFSS website (www.rrfss.ca) or see the attached list in **Appendix A**.

Health Unit ID

Another helpful thing to know about the data set is that each health unit is identified by a health unit ID. For example, the health unit ID for the Windsor-Essex County Health Unit is 20. The variable name for health unit ID in the data set is “h_unit”. You can find out the ID for your health unit either by looking in the variable list in the data file or by taking a look in the Questionnaire Map and the Technical Document.

Knowing your health unit ID is important since you receive data for all health units and will often want to select out the data for your area only. In order to do so, you need to be able to select only that data from your area. You do this using the health unit ID variable.

Weights

One final set of variables contained in the data set that are very helpful are the weight variables. Because RRFSS only samples a proportion of the population but the data is used to make statements about specific populations of interest, weights are required to ensure that the data is representative of those populations. For example, weights are needed when you want to make statements about the population in your health unit region based on the sample of RRFSS data.

RRFSS provides household weights to correct for bias based on the size of the household. These are required only for questions pertaining to individuals (e.g., self-reported chronic disease). For questions that are asked about the household (e.g., household income) no weighting is necessary.

Because respondents from Health Unit areas with larger populations have a lower chance of being interviewed than respondents from smaller Health Unit areas, the use of population weights to compensate for the unequal probability of selection at the Health Unit level is recommended.

While there is some debate as to the need for more complex weighting procedures, household weighting is generally agreed to be sufficient for most RRFSS analyses. For further detail on the nature of the weight variables provided and when to use them, see the section on **Using Weights**.

ISR-Derived Variables

Starting January 2007, ISR created some variables so they do not need to be created by the user. The derived variables are denoted with the suffix (_isr). So far the derived variables are:

- Age of respondent (age5_isr, age19_isr, age24_isr, age34_isr)

- Age of children in the household (nage0, nage0_6, nage1_3, nage1_4, nage4_11, nage5_9, nage5_11, nage6_11, nage12_17)
- Education (ed_ISR)
- Income (inc30_isr, inc40_isr, inc50_isr)
- Language (hlangrec_isr, mtongrec_isr)
- Fruits and Vegetables (fvtrunc_isr, fruitveg2_isr, fruitveg3_isr)
- BMI (bmicat_isr)
- Physical Activity (pa_level_isr)
- Alcohol Consumption (drnkday_isr, drnkwk_isr, lrdg_isr)
- Pregnancy Status (preg_isr)
- Smoking Status (smoke1_isr, smoke2_isr, sfree_isr)
- Bike Helmets (wearhel_isr)
- Booster Seats (boost_isr, boost2_isr)

Documentation for these is to be completed by ISR.
Please note that this list will expand as more variables are derived.

Analysis

Now that you have an understanding of the data set you may be eager to jump in and start analyzing data. In order to do so, RRFSS provides you with a number of indispensable tools. Two of the most useful tools are the RRFSS data dictionaries and existing syntax files.

Data Dictionaries

Individual RRFSS variables are grouped in modules. For every module used on RRFSS a data dictionary exists. Data dictionaries can be found most easily by going to the RRFSS website (www.rrfss.ca) and clicking on data dictionaries on the home page. For ones that are not posted on the website, please contact the RRFSS coordinator. This will provide you with a full list of links to most, if not all data dictionaries for existing modules. All data dictionaries have the same format and include amongst other things:

- Module title
- Module information
- Purpose of the module
- Module history
- Data dictionary history
- Module questions
- Variable names
- Variable history
- Question and response options
- Comparability to other surveys
- Validity/Reliability tests
- Evaluation questions
- Module indicators
- Method of calculation

Although all of these components are important to data analysis, the section most directly involved with analysis is the one dealing with module indicators and methods of calculation. Here you will find information on appropriate numerators and denominators, and any information relevant to performing calculations and data analysis. Note that for a small number of indicators, these calculations have been completed by ISR in the form of derived variables and are included in the dataset – see section on derived variables.

Syntax Files

Also extremely helpful to the data analysis process are existing syntax files. Over the years, individual RRFSS representatives and the Analysis Group have created syntax files (SPSS, Stata, SAS) to assist with the analysis of existing modules.

Typically, syntax can be found readily for core modules and to perform common types of functions (e.g., calculating confidence intervals). Where optional modules are concerned, typically those health units asking a particular optional module will have developed the necessary syntax and are usually happy to share.

Currently there is no central system for syntax development. Your best bet is to send an e-mail to all RRFSS representatives asking if anyone has developed a syntax for the module that you are interested in. But please note that at this time, you will be using the syntax at your own risk.

Using Weights

The RRFSS data set provides derived weights for a number of different data configurations for each health unit and across health units. The table below gives you a quick overview of the different sets of weights provided to you.

Data Set	**Weight Variable to be Used**	**Reminders**
By health unit for a specific cycle	*hhuwXX* – where XX refers to wave being analyzed	Select data for health unit and cycle first
By health unit for entire data set	*hhuall*	Select data for health unit first
All health units for a specific cycle	*hhwgtwXX* – where XX refers to wave being analyzed	Select data for cycle first
All health units for entire data set	*hhwgtall*	No selection necessary, use data for entire data set

These can all be a bit confusing at first. That said, the following examples may help:

> Example 1: You wish to analyze data specific to your health unit for cycle 19 (January – April 2015). First, you would select the data for your health unit (using variable H_UNIT) and then only select cycle 19. Then you would use the weight variable *hhuc19* to weight data prior to analyzing the data for your health unit for cycle 19.
>
> Example 2: You wish to analyze data for your health unit for the entire year of 2015. First ensure that you are using the full data set for 2015. As in the previous example you would first select the data for your health unit only. Unlike the example above, however, you don't need to select a specific cycle since you're using the complete data set. Then you can invoke the weight variable *hhuall* prior to conducting your analyses. This variable provides the appropriate weight for your health unit for the entire data set.

Conditions for Using Derived Weights

As previously stated, the *hhuall* and *hhwgtall* variables allow you to weight data for your health unit and all health units, respectively, for an entire data set. These weights are cumulative, meaning that for any given data set received over the course of a year, the *hhuall* and *hhwgtall* variables will be appropriate to use for the complete analysis for that entire data set. Specifically, any given data set to be weighted using these variables must begin with the first data collected (e.g., all of January) and contain all data for the remainder of the data set (e.g., Jan-Apr, Jan-Aug, Jan-Dec). Take this example:

> Example 3a: The 2015 January – August data set includes data from Cycle 19 (January-April 2015) and Cycle 20 (May-August 2015) along with the weights suitable for that data set. If you wanted to analyze the full data set for your health unit (from January to August), you would select your health unit's data and utilize the *hhuall* variable for weighting purposes.

Example 3b: The final 2015 data set includes data from Cycle 19 (January-April 2015), Cycle 20 (May-August 2015), and Cycle 21 (September-December 2015) along with the weights suitable for that data set. If you wanted to analyze the full data set for your health unit (from January to December), you would select your health unit's data and utilize the *hhuall* variable for weighting purposes.

Weighting for Partial or Combined Data Sets

While the derived weight variables in the RRFSS data set are quite helpful when analyzing a single Cycle from a partial or complete data set, they cannot be used when you wish to analyze a sub-sample of data from the middle of a data set or when you combine data sets across calendar years. There are a number of situations where these might be the case, for example:

- Seasonal Modules – Some variables are only asked in RRFSS in the middle of the year. For example, sun safety questions have traditionally been asked only from May to September.

- Optional Modules – Many health units choose to ask questions about a specific campaign they are running. Often these questions only run in the part of the year that corresponds to when their local campaign is active.

- Multi-Year Analyses – Particularly for modules that typically capture a small proportion of the overall sample (e.g., reproductive health – currently pregnant), it is useful to combine across years for enhanced data quality. This is especially true if you wish to look for specific characteristics of that sub-population (e.g., smoking status among women currently pregnant). Alternately, you may want a larger data for any number of reasons.

The question then becomes one of how to weight data when this is the case? The short answer is that you need to create your own weights in these instances. A number of different options are available to you. One way to address this issue is as follows:

- First, select the Cycles that you want. Then run a frequency on the number of adults in the household (variable is called "nadults"). From there, you calculate the total number of weighted cases by multiplying the number of cases (frequency) by the number of adults in the household. See example below:

Number of Adults in Household (nadults)	Frequency	Weighted Cases
1	322	322
2	691	1382
3	127	381

4	50	200
5	10	50
6	1	6
7	1	7
8	1	8
Total	1203	2356

- Then you use the syntax below to calculate the weight for the example above.

SPSS:

```
COMPUTE weight = nadults*1203 / (1*322+2*691+3*127+4*50+5*10+5*10+6*1+7*1+8*1).
FORMAT weight (f7.5).
EXECUTE.
```

Stata:

```
egen sumadult = total(nadults)
gen sumcase = _N
gen weight = nadults/sumadult*sumcase
```

For more details on weights and weighting, take a look in section **2.6 Household Size Weights** in the **RRFSS Technical Documentation**.

Confidence Intervals

A vital part of data analysis and reporting is the capacity to make statements about statistically significant differences. The following should be helpful:

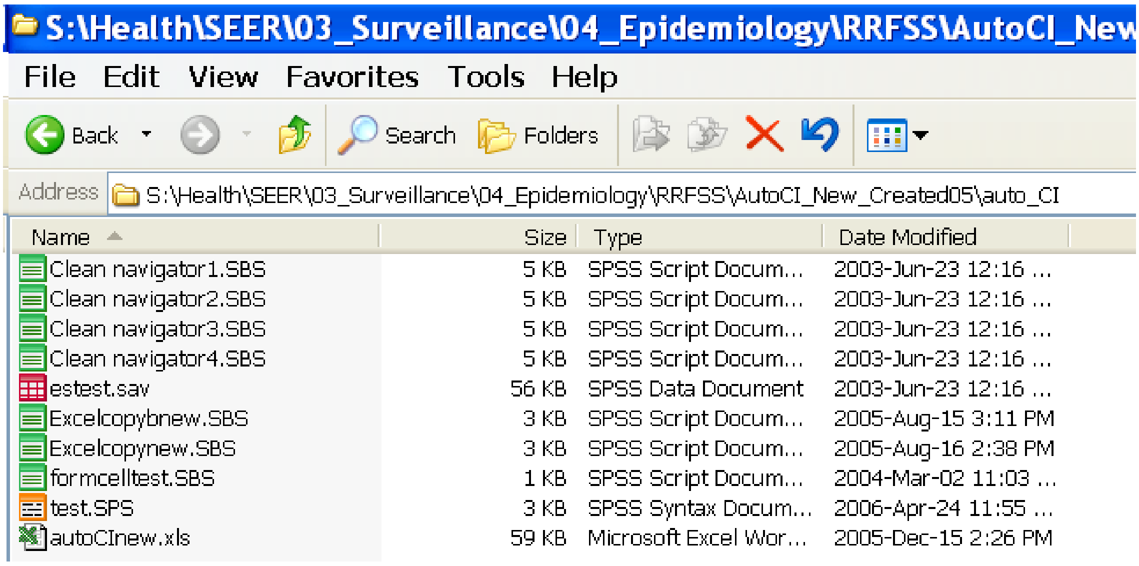
- Calculating Confidence Intervals (CI): As stated in the Analysis Guidelines in the MOO, the simple computation of the C.I. for a proportion assuming SEp = sqrt(pq/n) and CI95% = p +/- 1.96*SEp is sufficient. However, if estimates are close to 0 or 100% it is possible that the calculated CIs will reach below 0% or above 100%, which is non-sensical. In such cases, a CI calculation for skewed estimates (i.e., a logit transformation) should be used. This is often the default in modern software packages, but should be confirmed (Refer to Fleiss et al., 2003 [1] for more information). This approach will result in CIs being asymmetric about the estimate, which is appropriate.

[1] Fleiss, J.L., Levin, B., and Paik, M.C. (2003). *Statistical Methods for Rates and Proportion, 3rd Editions.* New York, NY: Wiley-Interscience.

Some health units have developed programs to calculate the CI, here are two examples:

- Halton Region Health Department developed a combination of a dataset and an accompanying syntax file that calculates the CI. To use this, all you need to do is to enter the following on the dataset called “ci.sav”: your estimate/proportion (p), the inverse of that (q) and your total (n). Then you run the syntax labeled “ci.sps” and the upper and lower CI, the CV will be calculated for you. You can obtain these two files from the RRFSS Coordinator.

- Another tool that is available to you is the “Auto CI” program developed by Bernie Leuske from London Middlesex Health Unit. To obtain copies of this program please contact the RRFSS coordinator. This is a program that uses a combination of SPSS script files and Excel macros not only to calculate the CI, but also to put all your output into tables and graphs. The Auto CI folder contains 7 SPSS script files, a test syntax, a test dataset and an Excel file, as shown below. For details on how to use this program, refer to **Appendix 2: Directions for Using the Auto CI Program**.



S:\Health\SEER\03_Surveillance\04_Epidemiology\RRFSS\AutoCI_New

S:\Health\SEER\03_Surveillance\04_Epidemiology\RRFSS\AutoCI_New_Created05\auto_CI

Name	Size	Type	Date Modified
Clean navigator1.SBS	5 KB	SPSS Script Docum...	2003-Jun-23 12:16 ...
Clean navigator2.SBS	5 KB	SPSS Script Docum...	2003-Jun-23 12:16 ...
Clean navigator3.SBS	5 KB	SPSS Script Docum...	2003-Jun-23 12:16 ...
Clean navigator4.SBS	5 KB	SPSS Script Docum...	2003-Jun-23 12:16 ...
estest.sav	56 KB	SPSS Data Document	2003-Jun-23 12:16 ...
Excelcopybnew.SBS	3 KB	SPSS Script Docum...	2005-Aug-15 3:11 PM
Excelcopynew.SBS	3 KB	SPSS Script Docum...	2005-Aug-16 2:38 PM
formcelltest.SBS	1 KB	SPSS Script Docum...	2004-Mar-02 11:03 ...
test.SPS	3 KB	SPSS Syntax Docum...	2006-Apr-24 11:55 ...
autoCInew.xls	59 KB	Microsoft Excel Wor...	2005-Dec-15 2:26 PM

Combining Data Sets

It is important to note that when combining datasets from different years, you must match on a number of variables: ID, Cycle/Wave, Health unit (if you are using more than one). The key message is that the Ids are not always unique and as such adding the wave is very important.

When combining datasets (stacking or appending files) you will need common variables to continue with an analysis. It is possible that response options and/or skip patterns may have changed and also possible that additional questions have been added or removed. For example, the derived variable **new_agegroup_isr** does not exist within the RRFSS dataset prior to 2014. Therefore, this variable cannot be used if analyzing datasets that predate its inclusion in the dataset. Similar scenarios will exist for various variables (e.g., Waves from 2005-2008, but Cycles from 2009-present, etc.

Appendix A: RRFSS Waves and Cycles by Month and Year

		MONTH											
YEAR		**January**	**February**	**March**	**April**	**May**	**June**	**July**	**August**	**September**	**October**	**November**	**December**
2001	**WAVES**	1	2	3	4	5	6	7	8	9	10	11	12
2002		13	14	15	16	17	18	19	20	21	22	23	24
2003		25	26	27	28	29	30	31	32	33	34	35	36
2004		37	38	39	40	41	42	43	44	45	46	47	48
2005		49	50	51	52	53	54	55	56	57	58	59	60
2006		61	62	63	64	65	66	67	68	69	70	71	72
2007		73	74	75	76	77	78	79	80	81	82	83	84
2008		85	86	87	88	89	90	91	92	93	94	95	96
2009	**CYCLES**	1				2				3			
2010		4				5				6			
2011		7				8				9			
2012		10				11				12			
2013		13				14				15			
2014		16				17				18			
2015		19				20				21			
2016		22				23				24			

Appendix C: Data Analysis General Principles and Guidelines, 2015 MOO revision.

The following are general principles and guidelines for the analysis of RRFSS data. They are based on the knowledge, experience and technical expertise of the RRFSS Analysis Group, as well as feedback from RRFSS Representatives. Their purpose is to facilitate consistency and hence comparability of the data analysis methods used by RRFSS-Participating Health Units.

While what follows will provide guidance applicable to most data analyses done by RRFSS-participating health units, other issues not covered here may occasionally arise. It is strongly recommended that all RRFSS-Participating Health Units pay careful attention to their results and determine if any health unit-specific issues exist with their data.

1. Data Release Guidelines

The survey analyst should examine the <u>unweighted</u> counts to determine if that estimate is suitable for public release. Estimated proportions based on RRFSS data shall be <u>suppressed</u> if either of the following conditions are TRUE:

i. a. The unweighted denominator of the proportion is less than 30 respondents; OR
ii. b. The unweighted numerator of the proportion is less than 5 respondents

Further guidance regarding data release can be provided by the coefficient of variation (CV). The CV should be calculated for each estimate using methods appropriate for the analysis of complex surveys – for example, by using the SPSS Complex Samples module or Stata's SVY commands.

Estimated proportions can be released without qualification if the CV is between 0 and 16.5. The estimate should be considered for <u>release with qualification</u> ("*Interpret with Caution: High Variability*") if the CV is greater than 16.6.

Further, the analyst should consider suppressing the estimate if the CV is greater than 33.3. Note that estimated proportions that are smaller are more likely to exceed this CV threshold, and may be considered for release if the unweighted counts for both the numerator and denominator are sufficient (see above).

2. Analysis of RRFSS as a Complex Survey

RRFSS employs a two-stage (i.e., 1-Household; 2-Respondent) sampling design. It is not a simple random sample. Analysis of RRFSS data should employ methods appropriate for the analysis of complex surveys, which are available in modern day software packages – for example, the SPSS Complex Samples Module or Stata's SVY commands. Refer to Heeringa et al., 2010 [2] for an excellent resource on methods for the analysis of complex surveys.

3. Estimation of Standard Errors / Confidence Intervals

As a survey based on a sample of the population, all estimates derived on RRFSS data will have some degree of sampling variability. All released RRFSS estimates should acknowledge and measure this variability by including standards errors and/or 95% confidence intervals for the estimate, based on the unweighted counts (n).

If estimated proportions are close to 0% or 100%, it is possible that the calculated CIs will reach below 0% or above 100%, which is non-sensical. In such cases, a CI calculation for skewed estimates (i.e., a logit transformation) should be used. This is often the default in modern software packages. Note that it will result in the CIs being asymmetric about the estimate, which is appropriate. Refer to Fleiss et al., 2003 [3] for more information.

4. Weighting

i) The RRFSS dataset includes a series of variables providing household weights. These weights compensate for the different probability of the respondent having been selected, depending on the number of adults in the household

The household weight should be applied during data analysis if:

a) the question is about the individual respondent; OR
b) the question is about the household, but the analyst wishes to determine the proportion of the population that lives in households sharing that particular characteristic (e.g., the % of the population living in a household using well water that has not been tested for bacteria in the past 12 months).

The process to calculate the household weight is:

- Total the number of adults (egen sumadult = total(nadults))

- Divide the row's nadults by the product of the total number of adults surveyed and the number of surveys completed (gen wt = nadults/sumadult*_N)

The household weight should <u>not</u> be applied for child proxy questions (bicycle helmet use, car seat safety), dog and cat immunization modules, or other questions that relate to the household rather than the respondent (e.g., the % of <u>households</u> using well water that has not been tested for bacteria in the past 12 months.

[2] Heeringa, S.G., West, B.T., and Berglung, P.A. (2010). *Applied Survey Data Analysis.* Boca Raton, FL: Chapman & Hall/CRC

[3] Fleiss, J.L., Levin, B., and Paik, M.C. (2003). *Statistical Methods for Rates and Proportion, 3rd Editions.* New York, NY: Wiley-Interscience.

Household weights are not required to be recalculated for sub-population based questions; for example mammography in women ages 35+ years and 50-74 years.

ii) If the weights supplied with the data set (health unit wave/cycle specific, health unit cumulative total, all health units combined wave/cycle specific, all health units combined cumulative total) are not appropriate for the required analysis, then a time-specific weight must be calculated. For example, a new weight is required for all seasonal modules.

Seasonal modules (modules ask for a 1 or 2 waves/cycles, not the whole year), do not use the same weight as modules asked through the entire 12 month period. If a seasonal module is asked for one cycle, then health unit wave/cycle specific weight (HHUC#) is to be used. If a seasonal module is asked for 2 cycles (which may or may not be sequential) then a time-specific weight is to be calculated. To calculate the seasonal weight, select the responses which took place in the wave/cycle of interest, or selected based on the interview date variable, depending on the starting and ending dates of the module and use the calculation from 4.1.

iii) If multiple years of RRFSS data are to be combined for an analysis, the household weight will need to be recalculated. After the dataset have been combined (appended), the weight for the cumulative file can be created using the formula in 4.1.

5. Basic Categorizations for Subpopulation Analysis

Unless an indicator requires a specific age grouping, age categorizations should be consistent with the following: 18-24, 25-44, 45-64, 65+. Income and education categories will be health unit-specific. The number of categories selected will be highly dependent on the cell size. For small numbers, categories may need to be collapsed in order to ensure valid and releasable results.

6. Inclusion of Non-Response Categories

Non-response categories (i.e., ‘Don’t Know’, “Refusal”) should be included in the denominator and reported separately if they exceed 5% of the all responses for a group.

7. Analysis of Data with Skip Patterns

When calculating indicators, the denominator used should be explicitly stated. This is of particular importance when there are skip patterns in the questionnaire and the denominator used is “*all adults who were asked the question*” as opposed to “*all adults*”.

8. Text Files and Post Survey Coding

Some survey questions include a response category where the respondent is asked to specify, usually “Other, specify”. Responses are typed out by the interviewer, saved as a text file and distributed with the data set for each cycle. Analysis of questions with this type of response should include a review of the open-ended responses. Health Units are responsible for reviewing their health unit-specific open-ended responses. In some cases, post-survey coding of the responses into new or existing response categories is required for accurate analysis of the data. Post survey coding is not an ISR responsibility.

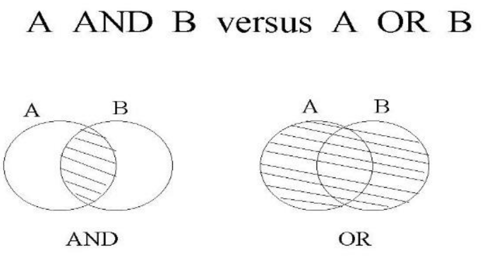
For post survey coding to be included in the final annual RRFSS data set produced by ISR the coding must be carried out for all health units that asked the question and for all waves/cycles of data. A lead health unit should be identified for the recoding project and the recoding completed by RRFSS representatives. The recoded data is forwarded to ISR and merged with the RRFSS data when the final data set is produced. If health units would like ISR to recode the data, a Special Request must be submitted and a cost estimate will be provided by ISR.

When the recoding is completed, the lead health unit will send the criteria used for recoding to the RRFSS Coordinator to include in the Analysis Issues of the Data Dictionary for the question recoded. When reporting results of questions where the text files have been recoded, the post survey coding should be noted.

If recoding is not done for all health units, individual health units can recode their own data. If individual health units would like ISR to recode their data this must be negotiated with ISR outside the RRFSS contract. Recoded data for individual health units will not be included in the RRFSS data set.

9. Boolean Operators

The meaning of ‘AND’ (&) and ‘OR’ (|) within a syntax file and/or the indicator calculation in the data dictionary pages are shown in the diagram below.



10. Syntax File Authorship

During syntax file development it is recommended that the name of the developer, the date of file development and a brief description of the analysis be included. Upon revision of syntax files it is recommended that the name of the reviser, the date of the revision and a brief description of the revision be included (see *Section 2.7 Procedure for Syntax File Sharing and Posting).*

Appendix B: Directions for Using the Auto CI Program (SPSS)

This program does not formulate your analysis for you. It is assumed that the proper weights and variables are being applied by the user. It is intended for use with RRFSS, and creates **unweighted tables** and **weighted tables** so that you may check N values (should be > 29) and cell counts (should be >4) in the **unweighted table**. The macro is only designed to work with the Crosstab procedure output. The following is important to know when navigating the Excel tables produced by the Auto CI program:

- If "Total" values are coloured blue in the weighted table then they are less than 30 in the unweighted table.
- If cell counts are less than 5, they are coloured red.
- If a C.V. is between 16.6 and 33.3 (release with qualification) the value will be colored green.
- If a C.V. is greater than 33.3 (should not be released) then it will be colored red.

Example: Smoking Status

This example uses smoking status as an indicator. The first step is to derive your indicator, for example daily, former, and never smokers, for your dataset. Then, proceed as follows:

1. Open excel spreadsheet "autoCInew.xls" (Select "enable macros" if prompted)

2. Open SPSS and open the estest.sav dataset. Open the test syntax file "test.SPS". Run this program and wait until SPSS displays the message "SPSS for Windows Processor is ready". Go into the Excel spreadsheet and make sure the output in the Paste worksheet matches the output in the Test worksheet. Basically look at the tables and see if the totals match and that two sets of crosstabs have run. After you have confirmed that the test worked, then clear the Paste worksheet and delete the charts or delete the worksheet and create a new Paste worksheet.

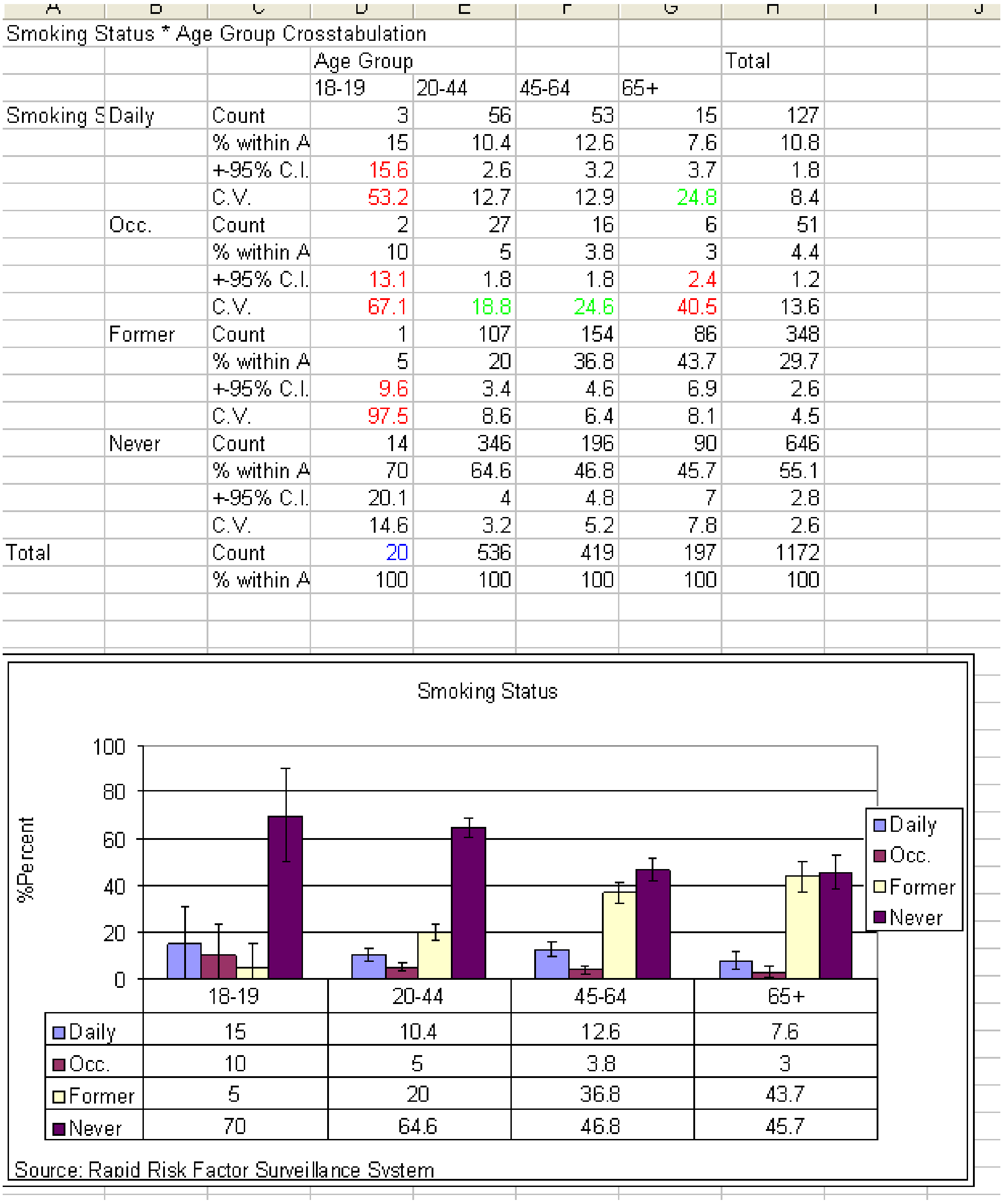
   Please note that after you become familiar with the program, you don't need to do this test again. You can just use your dataset right away and skip this part.

3. If the test has worked then go back to SPSS and open your data file. Select the syntax file "test.SPS" and modify this file with your crosstab variables and run it. It is important that you use the whole syntax file. The only parts that you need to modify are the crosstab variables and the weight if you are using a different weight. All the script files (clean navigator 1 through 4, the excelcopy.sbs and the formcelltest.sbs files) that are included in the syntax have to be used as is. You can run multiple crosstabs by modifying the tables portion for the crosstab command (e.g., /TABLES= es1 es2 by inc40). Again wait until SPSS displays the message "SPSS for Windows Processor is ready". Your Weighted Tables should have confidence intervals and C.V.'s with a basic Chart beside it.

4. You can rename the spreadsheet anything you want to match your project. (e.g., rrfssseatsmart.xls) and save it anywhere you want.

5. After running crosstabs you can just rename the "Paste" worksheet with your tables and charts something else and then insert a new worksheet and call your new worksheet "Paste". There always has to be a blank "Paste" worksheet for the program to work.

See the next page for a sample of the output of the cross tabulation of smoking status by age group. As you can see, you get the proportion, CI and CV for each age group and a chart with error bars. As explained above, when the colour of the CI is red it means that the CV is more than 33.3 and the estimate cannot be released. If it is green it is released with qualification.



Smoking Status * Age Group Crosstabulation							
			Age Group				Total
			18-19	20-44	45-64	65+	
Smoking S	Daily	Count	3	56	53	15	127
		% within A	15	10.4	12.6	7.6	10.8
		+-95% C.I.	15.6	2.6	3.2	3.7	1.8
		C.V.	53.2	12.7	12.9	24.8	8.4
	Occ.	Count	2	27	16	6	51
		% within A	10	5	3.8	3	4.4
		+-95% C.I.	13.1	1.8	1.8	2.4	1.2
		C.V.	67.1	18.8	24.6	40.5	13.6
	Former	Count	1	107	154	86	348
		% within A	5	20	36.8	43.7	29.7
		+-95% C.I.	9.6	3.4	4.6	6.9	2.6
		C.V.	97.5	8.6	6.4	8.1	4.5
	Never	Count	14	346	196	90	646
		% within A	70	64.6	46.8	45.7	55.1
		+-95% C.I.	20.1	4	4.8	7	2.8
		C.V.	14.6	3.2	5.2	7.8	2.6
Total		Count	20	536	419	197	1172
		% within A	100	100	100	100	100

Smoking Status

	18-19	20-44	45-64	65+
🟦 Daily	15	10.4	12.6	7.6
🟥 Occ.	10	5	3.8	3
🟨 Former	5	20	36.8	43.7
🟪 Never	70	64.6	46.8	45.7

Source: Rapid Risk Factor Surveillance System

Glossary of Terms

CATI	This stands for **Computer Assisted Telephone Interviewing** and is the software application that ISR uses to conduct RRFSS surveys. Interviewers follow the script provided by the CATI application and surveys flow according to the answers provided and predetermined skip patterns.
CI	**CI** stands for **confidence interval**. All RRFSS data should be released with a 95% confidence interval. This tells the reader the range within which a true estimate falls 95 times out of 100. You may also hear about something called **Auto CI**. This is a program developed by Bernie Lueske that connects SPSS and Excel to provide weighted and unweighted estimates, Excel tables, and confidence intervals.
Cycle	Since Jan 2009, each four (4) month period of data collection is referred to as a **cycle**. There are three (3) waves in a year of data. Modules are added and removed to the RRFSS survey by cycles.
CV	**CV** stands for **Coefficient of Variation** and is a measure of the variability of a particular estimate. RRFSS data reporting requirements allow estimates with a CV between 0 and 16.5 to be released without qualification. However, estimates with a CV between 16.6 and 33.3 can only be released with caution, while those estimates with a CV greater than 33.3 cannot be released.
DD	This stands for **Data Dictionary**. All RRFSS modules have a data dictionary that provides information about the variables in the module, their history, and feedback relevant to analysis.
Denominator	When calculating things like rates and proportions, the **Denominator** is the bottom number. This reflects the population you are analyzing. Sometimes this will be the entire population (i.e., adults 18 and over), while at other times it may be a specific subsection of the population (e.g., all those answering "yes" to a previous question). Knowing the appropriate denominator is crucial to correct analysis and reporting. As well, unweighted denominators with a cell size of less than 30 cannot be released.

Evaluation Questions	In order to determine whether RRFSS questions are credible, new questions are subjected to an evaluation process. The **Evaluation Questions** are asked of each proposed RRFSS question to evaluate it before accepting it. Results for Evaluation Questions can be found with the files that Anne Oram sends out every 6-8 weeks.
ISR	**ISR** stands for **Institute for Social Research**. Located at York University, ISR is responsible for all RRFSS data collection.
Modules	All RRFSS variables are grouped into **modules** according to their purpose (e.g., chronic disease). Modules can be either **core** or **optional**. Core modules are chosen by the RRFSS group and asked of all health units, while optional modules are chosen on an individual basis by participating health units.
MOO	This stands for **Manual of Operations**. The MOO contains all official guidelines and procedures for RRFSS.
Numerator	The **Numerator** is the top number in your data analysis. It reflects the characteristic that is being measured (e.g., number of people self-reporting an affirmative diabetes diagnosis). Note that numerator data with an unweighted cell size of less than 5 cannot be released.
UFQ	This refers to the **User Friendly Questionnaire** document which is done for each module. It contains the module questions and skip patterns only. They are created by ISR when a module is added to the CATI and then posted on the RRFSS website.
Q-Map	**Q-Map** is really just a short form for **Questionnaire Map** and like it sounds, provides a "map" of the RRFSS questionnaire. You will receive one Q-Map per month for each round of data collection and can use it to verify what modules are being asked by which health units.
Wave	Prior to Jan 2009, each period of data collection was referred to as a **wave**. There were 12 waves in a year of data. Modules were added and removed to the RRFSS survey by waves. Data collection began on the 11th of the month and ended on the 10th of the following month.
Weight	A **weight** is used to adjust data to provide a better estimate of a given population. ISR provides a certain number of weights to be used with RRFSS data that adjust for different sized households. Weights can also be calculated for sub-groups.

Contributors & Editors

2008
Amira Ali, Ottawa Public Health
John Barbaro, Simcoe Muskoka District Health Unit
Judy Brennan, Niagara, Regional - Public Health Department
Evelyn Crosse, Middlesex-London Health Unit
Julie Fraser, Windsor Essex County Health Unit
Michael King, Sudbury & District Health Unit

2016
Andrew Harris, Haliburton, Kawartha, Pine Ridge District Health Unit
Liza Mercier, Institute for Social Research
Lynne Russell, RRFSS Coordinator
Michael King, Sudbury & District Health Unit