

## Combining multiple health and demographic variables from *E-STAT* and *Health Indicators* for student projects

Joel Yan, Statistics Canada, [joel.yan@statcan.ca](mailto:joel.yan@statcan.ca), 1-800-465-1222

### Assignment Overview

Health issues are of great interest to students. Students can use the *Health Indicators* database on the Statistics Canada site, [www.statcan.ca](http://www.statcan.ca) and on E-STAT to look for possible relationships between demographic factors, and health outcomes, or to look for patterns in health data by subprovincial health regions. The *Health Indicators* database contains a wealth of free data, as well as related articles from Statistics Canada and the Canadian Institute for Health Information. The Census and the Canadian Community Health Survey are two of several sources used to produce these data. Health related data are provided for Canada, the provinces and territories, and the 130 health regions across Canada. By correlating variables across these areas, students can assess the likelihood that there is a relationship among the variables.

### Objectives:

This assignment has two purposes:

- to demonstrate the wealth of data available for student analysis on the Health Indicators database on E-STAT and how the data can be analyzed using software available to students;
- to teach the skills and detailed procedures for extracting health related datasets from E-STAT and importing it into spreadsheet software (e.g. Excel, Quattro Pro) or statistical software (e.g. Fathom) available in schools.

### Questions for analysis:

- Which demographic factors are correlated to smoking?
- Do high levels of being overweight correlate to high levels of diabetes?

### Related Expectations for the Ontario Mathematics of Data Management (MDM4U)

**Course** (with the corresponding MDM4U course unit and curriculum document page numbers shown in brackets):

- Solve problems involving complex relationships with the aid of diagrams. (ODV.02, *Overall expectations* – page 49)
- Locate data to answer questions of significance of personal interest by searching well-organized databases. (OD1.01, *Organizing data* – page 49)
- Describe the relationship between two variables by use the use of scatter graphs and interpreting the correlation coefficient. (ST4.01, and ST4.02, *Statistics* – page 52)

**Duration:** one 75 minute period for Analysis 1 and 2. Doing Analysis 3 will require another period or homework using excel or software available to students at home. Analysis 4 is designed for students doing a major project on a health topic. Students could also start with the ready-made Fathom or Excel datasets provided with this lesson.

## Using E-STAT and *Health Indicators* to Research Health Issues Student Worksheet

### *Access Health Indicators Data and E-STAT*

1. On the Statistics Canada website home page, [www.statcan.ca](http://www.statcan.ca), select your language of choice, and then click on '**Our products and services**' (in the top blue bar).
2. Choose '**Free**' under 'Browse our Internet publications'
3. Scroll down and click on the subject '**Health**' then choose '**Health Indicators**'. How many different editions of the 'Health indicators' product are available for free downloading from this web site? \_\_\_\_\_ (Answer: 8 as of April 2004)
4. Click on the latest release - **Volume 2004, No. 1 (June 2004)**. Scan the page and answer:

**Question:** How many health indicators does this product provide at the health region, province/territory, and Canada level? \_\_\_\_\_ (Answer: over 80)

Now you have many choices for selecting data. Click '**Data tables**' on the left side bar.

5. Data tables are organized here according to the Health Indicator framework:

[Health status](#)

[Non-medical determinants of health](#)

[Health system performance](#)

[Community and health system characteristics](#)

In this case we are looking for smokers by health region and want to see if there is any correlation with average educational attainment.

6. Click 'Non-medical determinants of Health' to go to the large menu of tables available. Select Smoking to see the list of tables on Health Indicators for this topic.

**Question:** How many Health Indicators tables are listed related to smoking? \_\_\_\_\_

7. On the Statistics Canada website home page, [www.statcan.ca](http://www.statcan.ca), select your language of choice, and then click on '**Learning Resources**', the second button on the left side bar.

8. Scroll down and click the E-STAT purple bar.

9. Scroll down and click 'Accept and enter' to gain entry to the E-STAT software. E-STAT contains a wealth of data on the health, economics and demographics of Canadians.

10. Select Search CANSIM on the left side bar. Select 'Keyword Search' and hit Continue.

11. Enter 'smoking', (the first health determinant that we will analyze) in the Search for box and click 'Find Tables'. **Question:** How many tables are listed related to smoking? \_\_\_\_\_

Note: This includes the Health Indicators tables and more.

12. This brings up a list of CANSIM tables containing data related to smoking. Select 'Table 105-0027 Smoking, by age group and sex, household population aged 12

and over, Canada, provinces, territories, health regions and peer groups' since these data are available by health region.

13. Under **Geography**, scroll down and select all the health regions in one province or region of interest. You select areas by highlighting them on the list. In the sample graphs shown later, we selected the 37 Public Health Units in Ontario. The Ontario Public Health Units come immediately after the list of District Health Councils in Ontario on the list of geographic areas. Start by selecting the Algoma Public Health Unit and include all the 37 Ontario public health units up to and including the Toronto Public Health Unit. Note: Public Health Units are defined in Ontario only and mostly roll up to larger District Health Councils.

**Question:** As indicated in brackets after the word **Geography**, for how many different geographic areas do we have the data on smoking rates? \_\_\_\_\_ (Answer: 199).

14. Under **Age group** select 'Total, 12 years and over'.

15. Under **Sex** select 'Both sexes'.

16. Under **Smoking**, click 'Select all' to select all 7 characteristics related to smoking habits.

17. Under **Characteristics**, select 'Percent'.

18. Now select 'Retrieve as a Table'.

19. On the next screen under **output format**, click on the drop-down box and select 'HTML Table, Geography as rows'. Click 'Go' at the bottom of the screen to output the data with one row of data for each health region.

20. Inspect the data in the resulting table and answer the questions below.

What does each column contain? \_\_\_\_\_

How many columns of actual numeric data are there? \_\_\_\_

How many rows of actual data are there? \_\_\_\_\_ (Answer: 37)

Does this correspond to the number of health regions you selected? \_\_\_\_ If not, go back and redo the Geography selection before continuing.

21. Using your mouse highlight all the contents of the HTML table, including the column labels, but not the footnotes. Indicate that you want to copy this by hitting Ctl-C or from the pull-down Edit menu selecting Copy.

22. Next open your spreadsheet program. Position your mouse at the first cell and Paste in the contents.

23. Clean up the attribute names by deleting the footnote references and shortening the health region names if you wish.

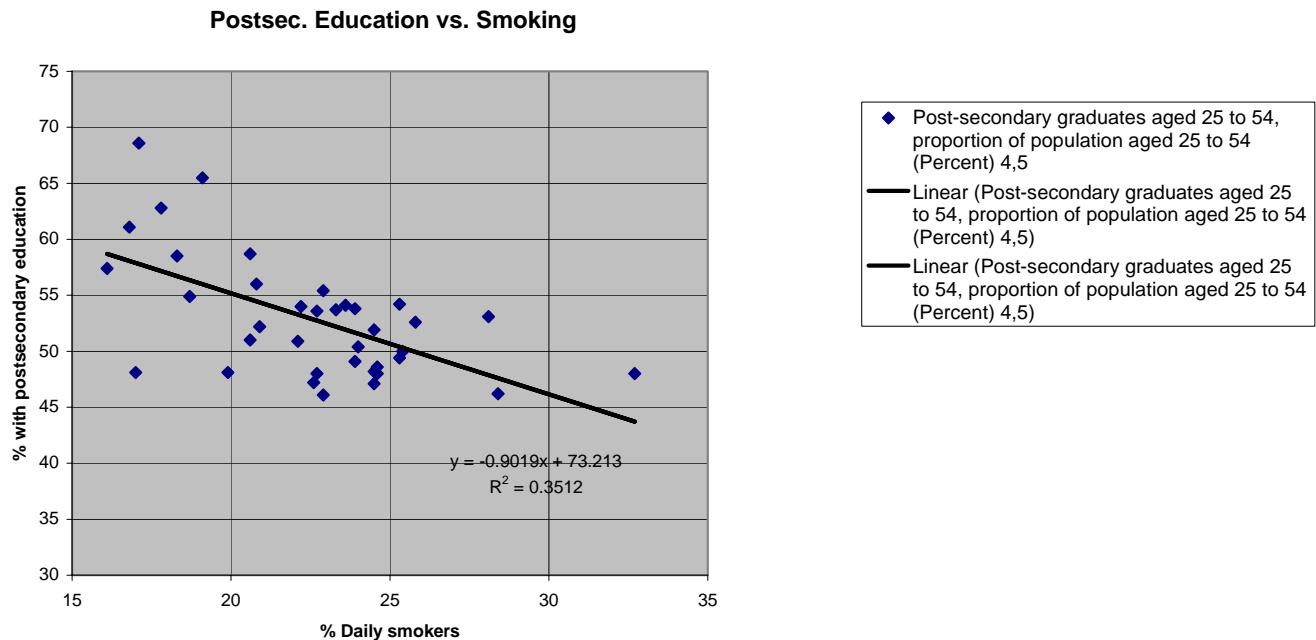
*Extracting the smoking data and demographic data into a spreadsheet*

24. Save the spreadsheet to disk with a file name “Smokers”. We will return to this file later to append possible explanatory or related factors.
25. Return to your E-STAT window. Now we will select demographic attributes for the same areas so that we can research potential relationships with health outcomes. Click 'Search CANSIM' in the left side bar. Click Search by 'Table Number ' and click Continue.
26. Enter 109-0200 in the box and then click 'Continue' to indicate you want to open this CANSIM table. This opens Table 109-0200, Census indicator profile, Canada, provinces, territories and health regions. This table contains several demographic variables, including education levels, for the same health regions.
27. Under **Geography** scroll down and select exactly the same regions that you selected above (e.g. Public Health Units for Ontario). This is required since we want to be able to compare and correlate some census variables with the smoking data already extracted.
28. **Questions:** How many characteristics are included in the Census Profile? \_\_\_\_ (47)  
Note: Each of these could be examined as a possible explanatory variable for the smoking rates or other health data we will be examining.  
Scan the list and name one characteristic on this list that you feel might be correlated to smoking rates by health region? \_\_\_\_\_
29. Under **Census profile** select the following and one other characteristic of interest:
  - High school graduates aged 25 to 29, proportion of population aged 25 to 29 (Percent)
  - Post-secondary graduates aged 25 to 54, proportion of population aged 25 to 54 (Percent)
  - Long-term unemployment rate, labour force aged 15 and over
30. Now select 'Retrieve as a Table'.
31. On the next screen under **output format**, click on the drop-down box and select 'HTML Table, Geography as rows'. Click 'Go' at the bottom of the screen to output the data. Ensure that the output corresponds exactly to the health regions you selected. If not, go back and redo the Geography selection before continuing.
32. Using your mouse highlight all the contents of the HTML table, including the labels. You do not need to include the footnotes and source notes in the domain to be copied. Indicate that you want to copy this by hitting CTL-C or selecting Copy from the pull-down Edit menu.
33. Open the spreadsheet file that you previously created for the smoking data. Position your cursor on the row where you have the attribute names in the row just above the first observation (this should be the data for the first selected health unit) and on the first blank column available. This should be just to the right of the smoking rates. Then Paste in the new data from the other spreadsheet. Tip: Use the Paste icon or select Paste from the 'Edit' pull-down menu.

34. Now check that the geographic areas from the two files match. Scroll down the spreadsheet and verify that the two area names on each row (from the two different CANSIM Tables) are properly aligned and correspond. If the first few names are not aligned properly, undo the paste that you just did, carefully position the cursor and redo the previous steps. Once you have verified that all the areas match, delete the second column containing the area name.

### *Graphing the data in Excel*

We can graph the data in Excel (as shown below) or import the data into Fathom and do



the analysis there.

### *Extracting other health variables of interest*

35. Return to your E-STAT window. Now we will select other health characteristics for the same areas. Click 'Search CANSIM' in the left side bar. Click Search by 'Table Number ' and click Continue.
36. Enter 105-0100 in the box and then click 'Continue' to indicate you want to open this CANSIM table. This opens up CANSIM table 105-0100, Canadian Community Health Survey (CCHS) indicator profile, by sex, Canada, provinces, territories and health regions. This table contains health characteristics for the same health regions.

37. Under **Geography** scroll down and select exactly the same regions that you selected above (e.g. Public Health Units for Ontario). This is required since we want to be able to compare and correlate some census variables with the smoking data already extracted, on the CANSIM screen for the selected table.
38. Under **Sex**, select Both sexes.
39. **Questions:** How many characteristics are included within the Health Profile? \_\_\_\_  
(33)
40. Under **Health profile**, click 'View checklist and footnotes'. Scroll down and select characteristics that you want to research. E-STAT will let you extract a maximum of 1000 data cells at a time. Since in this example shown we have selected 37 Ontario public health unit areas, we can extract a maximum of 27 health characteristics at one time. To begin with, select the first 26 characteristics on the list. Note: The elapsed time for the computer to process this step will be shorter if you select fewer characteristics
41. Click 'Return to pick list'
42. Under **Characteristics**, select Percent, so that easy comparisons can be made among health regions of different populations.
43. Now select 'Retrieve as a Table'.
44. On the next screen under **output format**, click on the drop-down box and select 'HTML Table, Geography as rows'. Click 'Go' at the bottom of the screen to output the data. This again outputs the data with one row of data for each health region. Ensure that the output corresponds exactly to the health regions you selected. If not, go back and redo the Geography selection before continuing.
45. Using your mouse highlight all the contents of the HTML table, including the labels. Indicate that you want to copy this by hitting CTL-C or selecting Copy from the pull-down Edit menu. Note: You do not need to include the footnotes and source notes in the domain to be copied

### ***Merging the health data with the existing Spreadsheet File***

46. Open the spreadsheet file that you previously created for the smoking data. Position your cursor on the row where you have the attribute names in the row just above the first observation (this should be the data for the first selected health unit) and on the first blank column available. This should be just to the right of the smoking rates. Then Paste in the new data from the other spreadsheet. Tip: Use the Paste icon or select Paste from the 'Edit' pull-down menu.
47. Now check that the geographic areas from the two files match. Scroll down the spreadsheet and verify that the two area names on each row (from the two different CANSIM Tables) are properly aligned and correspond. If the first few names are

not aligned properly, undo the paste that you just did, carefully position the cursor and redo the previous steps. Once you have verified that all the areas match, you can again delete the second column containing the area name.

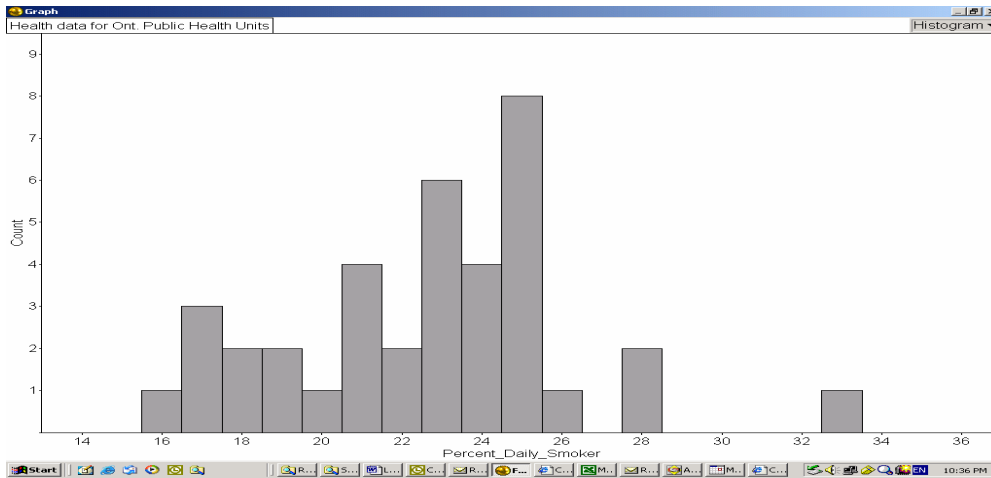
48. Scroll across the columns and delete any columns where the data are missing or blank as indicated by ..
49. Highlight the rows and columns of data in the spreadsheet and click the Copy icon. Choose **Copy** from the **Edit** menu.

### *Importing the Combined Data Variables into Fathom*

50. Switch to **Fathom**. (If Fathom isn't already running, you will need to launch it)
51. In a new document, make a new empty collection.
52. With the collection selected, chose **Paste Cases** from the **Edit** menu. This will import the health data from the spreadsheet.
53. Double click on the collection box. Change the name of the collection to a meaningful name, such as 'Health Data for Ont. Public Health Units'.
54. Make a case table for the collection (for example by choosing **Case table** from the **Insert** menu)
55. Edit the attribute names. Double click on each name in turn and shorten the names of the selected attributes as appropriate. For examples, change selected attribute names to Health\_District, Percent\_Daily\_Smoker, Percent\_HighSchool\_Graduates, and Percent\_Postsecondary\_Grad\_Aged25to54.
56. If the first case does not have numeric values for the two numeric attributes, delete this case. Scroll quickly through the case table. If there are any cases for which most numeric values are suppressed, delete these cases, by highlighting them, pulling down the Edit menu, and then selecting 'delete case'.
57. Save the Fathom document by choosing **Save** from the **File** menu.

### *Analysis 1 – Does smoking correlate to education levels?*

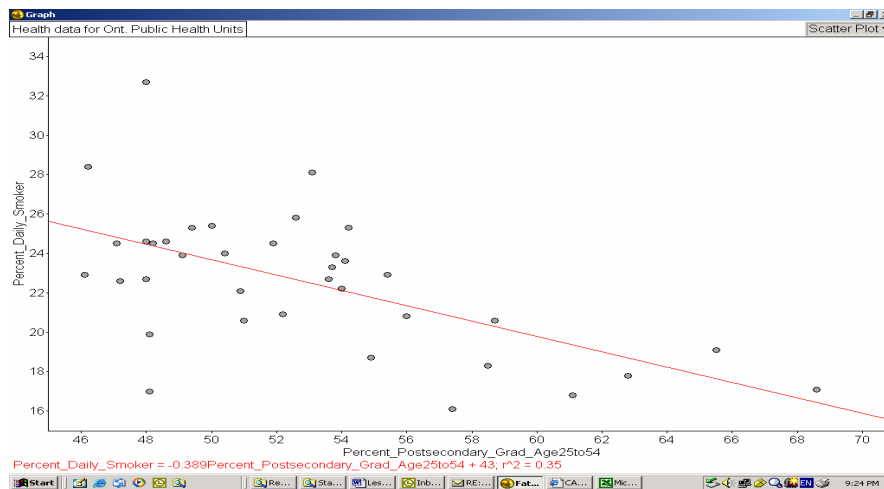
58. Bring down a graph. Drop the Percent\_Daily\_Smoker attribute on the x-axis. Hold down the CTL key as you do this, to exclude any of the non-numeric values from appearing on the graph. Using the pull-down menu above the graph, change the graph type to a histogram.



59. Write a description of the pattern you see on this graph. \_\_\_\_\_

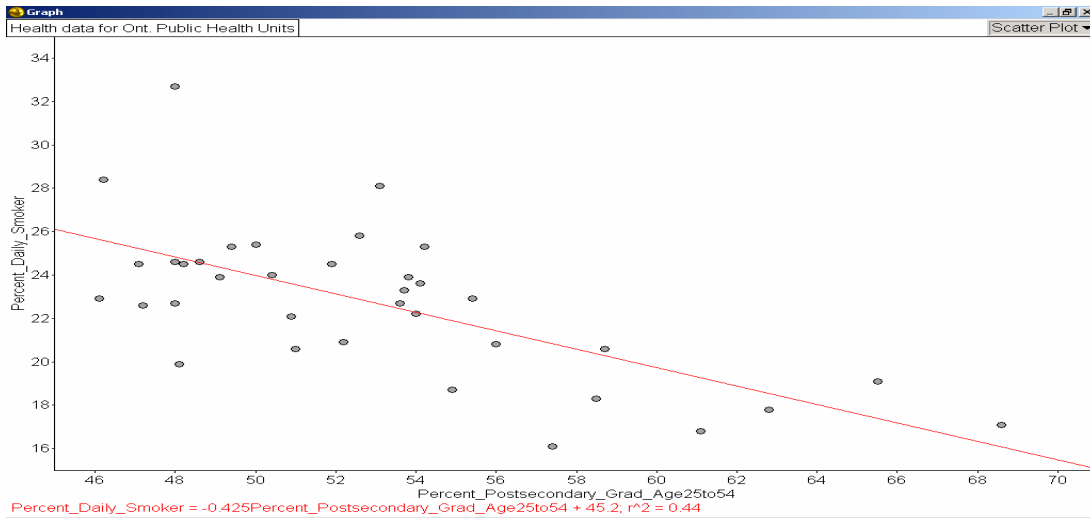
60. Bring down another graph. Drag the educational attainment attribute to the x-axis and the Percent\_Daily\_Smoker attribute to the y-axis

61. Right click on the graph and select the 'Least-squares Line' option. This will overlay in red the least- square line, as well as provide its equation and R squared value, as shown below.



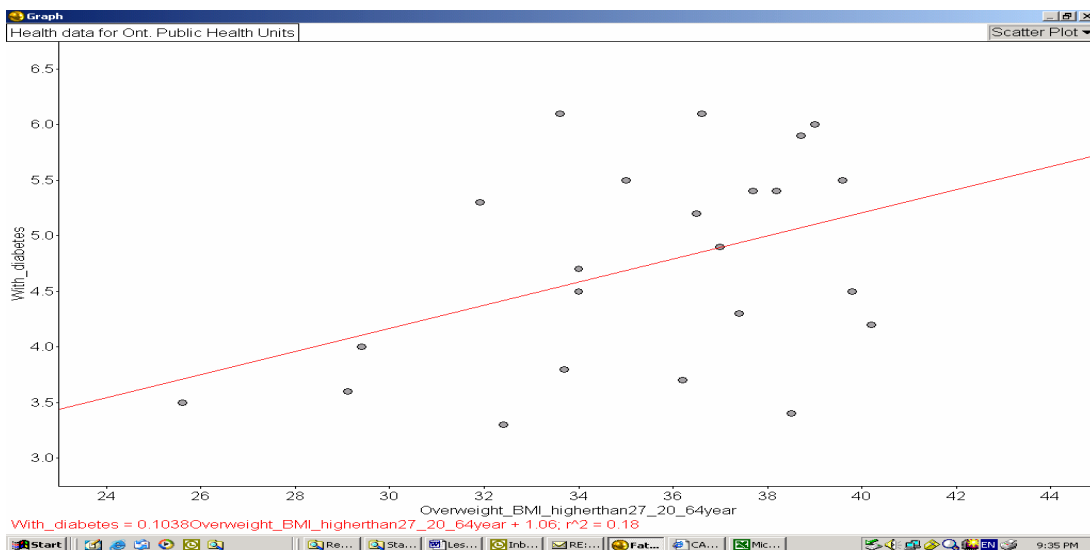
62. Analyze the outliers, those values that are furthest from the line of best fit and have the biggest impact. Highlight the farthest outlier and identify its location (this appears on the lower left of the screen).

63. **Questions:** Why might this area be an outlier? \_\_\_\_\_ Now click this outlier point, and delete this case. What is the impact on the  $r^2$  value and on the slope of the line? \_\_\_\_\_ (Answer:  $r^2$  changes from 0.35 to 0.44 and the slope changes from  $-0.389$  to  $-0.425$ ). After you have seen the impact on the equation, click Edit Undo to restore that point.



### ***Analysis 2 - Is there a connection between being overweight and having diabetes?***

Use this same Fathom collection of data by health district to explore the percent of the population that is overweight vs. the percent with diabetes. Below is the graph you can obtain. Note: You should hold down the CTL key as you do this, to exclude the non-numeric values from appearing on the graph.



### ***Analysis 3 – Find another relationship using this Fathom collection***

Repeat the steps shown in analysis 1 above to explore graphically possible relationships using other variables contained in the Fathom file.

Some relationships you may consider exploring to start with:

- chronic conditions such as diabetes, asthma, high blood pressure against any of the non- medical determinants of health;
- level of depression versus long term employment or education level.

### ***Analysis 4 - Find another relationship using other data from Health Indicators***

Select other variables related to Health status and Non-medical determinants of health. The **Health Indicators** product contains many such variables. E-STAT actually contains a number of other interesting CCHS profiles (e.g. urban- rural, aboriginal, immigrant) for provinces and territories. Data can be combined from many of these profiles with data from the CCHS Health profiles as indicated above as long as the geographic areas match. Again the variables can be combined and imported into Fathom, a spreadsheet, or other analytical software for further analysis.

Some relationships you may consider exploring:

- life expectancy against smoking, drinking, or life stress;
- lung cancer mortality against exposure to second hand smoke;
- infant mortality;
- circulatory disease death against level of physical activity, or against BMI;
- cancer deaths against dietary practices;
- deaths due to specific conditions against disease prevention measures such as flu shots, mammography, pap smears;
- level of depression versus income
- chronic conditions such as diabetes, asthma, high blood pressure against any of the other non- medical determinants of health.

**Caution:** For whichever of the more than 80 health indicators you extract for your analysis, remember to select exactly the same health regions (under Geography) and similar time periods. Keep in mind that some tables are not available at the health region level, and data from the Canadian Institute for Health Information are limited to regions with population greater than 75,000. Once the data have been merged in a single spreadsheet, check that all the geographic areas match.

File: Lesson4b –Combining Health Variables.doc

Folder: MDM4U/ Lessons

Last updated: May 20, 2004 based on subject matter review by Brenda Wannell

Related Excel Dataset: Lesson4b –Smoking, CCHS and demographic data for Ont. PHUs.xls

Related Fathom collection: Lesson4b –Smoking, CCHS and Demographic data for Ont. PHUs.ftm