TDU

Introduction to R For Public Health Investigations

Pre-Course Self-Study Module

Public Health Agency of Canada
Agence de la santé publique du Canada

Canada

Table of Contents

COURSE DESCRIPTION

DESCRIPTION

Introduction to R for Public Health Investigations is designed to help applied public health practitioners and field epidemiologists to develop the necessary skills to use R when conducting public health investigations. Skills are built gradually over the duration of the course; participants will use R in the context of outbreak investigations, public health surveillance, data management, and dissemination of information. Note: the intent of this course is not to provide comprehensive training in 1) outbreak investigation, public health surveillance, or scientific communication, 2) biostatistics, or 3) R overall. This course is oriented to provide epidemiologists and public health practitioners with additional skills and resources for use in the field.

This pre-course self study module was created to give participants the opportunity to review and familiarize themselves with the course pre-requisites at their own pace in advance of the course. The content of this module is split into essential pre-learning, required self-learning, and optional content, with the first two sections including content related to the virtual classroom training. The third section, optional content, includes materials that will be useful to learners who wish to go further on their R learning journey. Essential pre-learning items are for all learners to review with the assistance of self-assessment questions. These items are essential to learners' ability to participate in the course. Required self-learning items are particularly important for those:

- Who have never used R or who rely mainly on spreadsheet software (i.e., Excel) rather than statistical software (i.e., SAS, STATA, R) for analytical tasks;
- Who have used R in the past and would like to familiarise themselves or learn more on specific topics; or
- Who are proficient in R and looking for a refresher or to learn more on specific topics.

Appended to this document is the Pre-Course Checklist for the four day Introduction to R for Public Health Investigations course. Learners should review and complete each item listed.

LEARNING OBJECTIVES

By completing the pre-course content, participants will be able to:

- **Install** R, RStudio, and any packages required for performing desired analyses
- **Discuss** fundamentals of data analysis in R, including
    - Importing data from various formats
    - Principles of data management for public health
    - Dates and date formats in R
    - Visualizing results using ggplot
- **Identify** available resources for troubleshooting errors in R code and learning more about R packages

PRE-COURSE SELF-STUDY OVERVIEW

We have a lot of resources to share and don’t want you to feel overwhelmed! Throughout this document we will provide four different types of resources:

Mandatory Sea Time: Essential Pre-Learning

These items are essential to your participation in Introduction to R for Public Health Investigations. If these topics are new to you, please make sure to review the items prior to joining the course. This section includes information on how to succeed in virtual training with the Training and Development Unit, how to install R and RStudio, and how to install packages.

Swab The Decks: Required Self-Study

Please complete prior to attending the virtual training sessions. We have included self-study exercises in this document. These exercises are for provided for your use to assist in retention of content reviewed in self-study.

There Be Treasure: Going Further

Additional content provided as reference for self-learning and further development. Please note: this content is optional and not required to complete before participating in the course.

Appendix 1: Pre-Course Checklist

We have appended a Pre-Course Checklist to the end of this document. This checklist contains information to help you get ready for the course. We look forward to seeing you soon!

ESSENTIAL PRE-LEARNING

Before starting an ‘Introduction to using aRrr for public health’ course, you’re obviously going to need to have the proper software installed! Please complete the following tasks to make sure you’re up to date with all the tools you’ll need for the course! We’ve provided links to tutorials to walk through how to complete each of the steps in case you need a refresher or R is new to you.

Mandatory Sea Time: Essential Pre-Learning

The included essential pre-learning items will walk you through virtual training and technical requirements for starting Introduction to Public Health Investigations. See the descriptions and links below.

Have you:	**If not, please review:**
Watched “Virtual Training with the Training and Development Unit”?	☐ Video: Virtual Training with the Training and Development Unit
Watched “Finding Success with Virtual Training”?	☐ Video: Finding Success with Virtual Training
Installed R and RStudio?	☐ Tutorial: Install R and RStudio
Familiarized yourself with the RStudio interface?	☐ Video: Overview of RStudio User Interface
Learned how to install R Packages?	☐ Tutorial: Install R Packages
Ever imported data into R?	☐ Tutorial: Importing Data into R

How to Succeed in Virtual Training: Virtual Training with the Training and Development Unit

Duration: **5 minutes**

This video provides an overview of the Training and Development Unit's approach to training in the virtual environment: https://www.youtube.com/watch?v=lTDS6KgIGPg

How to Succeed in Virtual Training: Finding Success with Virtual Training

Duration: **4 minutes**

This video provides an overview of how learners can find success in virtual training with the Training and Development Unit: https://www.youtube.com/watch?v=CNo4rIg4nU8

Technical Check: Install R and RStudio

Learners should have R (4.2.2.0) and RStudio (2022.12.0.353) installed on their computers prior to joining the course. Those joining from CFEP will have this software pre-installed on their PHAC computers. Others joining from other areas within PHAC will need to open a ticket with the National Service Desk to have this software installed on their machines in advance. In the event individuals wish to join the course using computers not issued by PHAC, the following tutorial will aide them in installing R and RStudio on their computer themselves.
https://techvidvan.com/tutorials/install-r/#install-r-windows

Technical Check: Overview of RStudio User Interface

Duration: **6 minutes**

The RStudio interface makes R much more user friendly and allows for many different aspects of the program to be available at the user's fingertips through a series of panels. However, this setup still may be foreign and intimidating to the novice user. This YouTube video indicates what is available where: https://www.youtube.com/watch?v=5YmcEYTSN7k

Technical Check: Install R Packages

This tutorial reviews how to install packages in R. It is critical that learners understand how to install packages prior to joining the course. Note: this tutorial covers features that will not be required in the course (e.g., R-Forge, Bioconductor, Jupyter Notebook): https://r-coder.com/install-r-packages/

Technical Check: Importing Data into R

Importing data into R is the first step of any data analysis and is therefore important to be comfortable with prior to starting the course. This tutorial covers several different ways of loading data into R, including flat-format files (e.g., csv, txt), files from Excel (e.g., .xlsx, .xls), loading r-data files (.rds) and provides additional links for connecting and loading data sets from common database programs (e.g., MySQL). Learners for the Introduction to R for Public Health Investigations course should focus on loading flat files and files from Excel: https://uc-r.github.io/import

REQUIRED SELF-LEARNING

This wouldn't be much of a self-study module without providing some homework for you to study from! Please complete the following videos and tutorials that introduce the concepts that we'll be applying to public health topics in the course. Note: If you are an advanced R user, please review the descriptions for each of the Required Self-Learning items to see if it is something that will fill existing gaps in your knowledge.

Swab The Decks: Required Self-Study Checklist

Please review the descriptions (including links) below for each of the learning items listed here. Complete the following self-learning activities if the topic is one that you are unfamiliar with or where you need a refresher. At the end of this section we have included some self-study exercises. Please complete these exercises to consolidate and further develop your new skills and knowledge acquired through this self-study module, or to self-assess your pre-existing skills and knowledge.

Topic:

☐ Video: What is data wrangling?

☐ Video: Tidy data and tidyr

☐ Video: Data Manipulation Tools and Dplyr

☐ Video: Dates in R

☐ Video: ggplot

☐ Video: R Markdown

☐ Tutorial: Troubleshooting

☐ Video: Silly mistakes we all make in R/RStudio

☐ Webpage: Common R Errors

Foundational Concepts: What is Data Wrangling?

Duration: **9 minutes**

Data wrangling is too often the most time-consuming part of data science and applied statistics. Two tidyverse packages, tidyr and dplyr, help make data manipulation tasks easier. Keep your code clean and clear and reduce the cognitive load required for common but often complex data science tasks. This video is the first in a series of four, reviewing concepts such as tibbles, viewing data, the pipe operator, and data wrangling generally.
https://www.youtube.com/watch?v=jOd65mR1zfw

Foundational Concepts: Tidy data and tidyr

Duration: **18 minutes**

Data wrangling is too often the most time-consuming part of data science and applied statistics. Two tidyverse packages, tidyr and dplyr, help make data manipulation tasks easier. Keep your code clean and clear and reduce the cognitive load required for common but often complex data science tasks. This video is the second in a series of four, reviewing concepts such as tidy data and the tidyr package. https://youtu.be/1ELALQIO-yM

Foundational Concepts: Data Manipulation Tools and Dplyr

Duration: **20 minutes**

Data wrangling is too often the most time-consuming part of data science and applied statistics. Two tidyverse packages, tidyr and dplyr, help make data manipulation tasks easier. Keep your code clean and clear and reduce the cognitive load required for common but often complex data science tasks. This video is the third in a series of four, reviewing concepts such as the dplyr package, key functions for data manipulation, and how to chain statements together with the pipe operator. https://www.youtube.com/watch?v=Zc_ufg4uW4U

Foundational Concepts: Dates in R

Working with dates and times can be one of the most challenging parts of any statistical software too. The following tutorial covers the basics of date, time, and date-time classes in R, and is a good place to start to begin learning about these concepts, and why they can be so challenging: https://www.neonscience.org/resources/learning-hub/tutorials/dc-convert-date-time-posix-r

The **lubridate** package, written for the expanded tidyverse, was created to help alleviate some of the common frustrations in working with date and time data. Please review Chapter 10.3 of the following resource to gain an overview of common functions within the **lubridate** package:
https://bookdown.org/hneth/ds4psy/10-3-time-lubridate.html

Working with dates and times can be challenging and complex – so please don't feel that you need to master these concepts before joining the course. If you want to learn more about working with these data types – check out the additional resources included at the end of this pre-course self study guide!

Foundational Concepts: ggplot

Duration: **5 minutes**

This YouTube video reviews the use of the ggplot2 package to create a scatter plot and histogram. This video makes use of a pre-installed dataset allowing for viewers to follow along: https://www.youtube.com/watch?v=ccLi41JwkbQ

Foundational Concepts: R Markdown

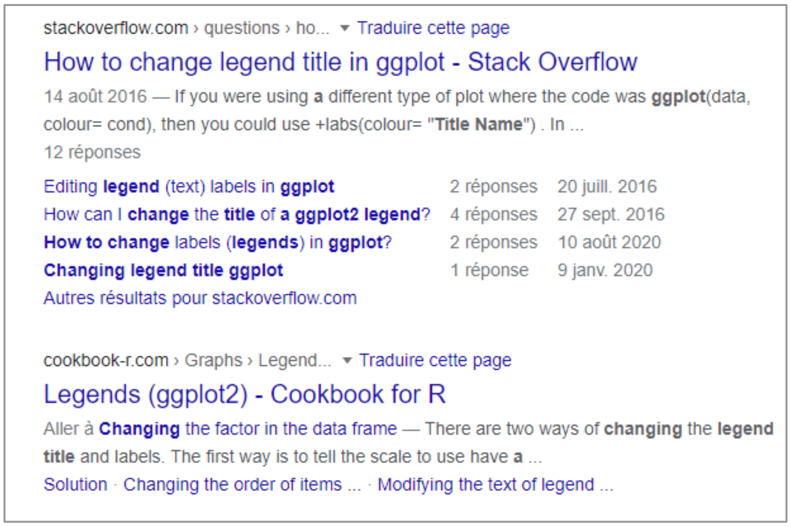
Duration: **7 minutes**

This YouTube video provides a demo of R Markdown: https://www.youtube.com/watch?v=DNS7i2m4sB0

Knowledge Boost: Troubleshooting

A few ways/resources for troubleshooting R issues:

- Every analyst has had the experience of hitting a wall when trying to code. What to do?
- R requires an almost detective-like attitude: the answer is out there online, you just need to sleuth it out.
- Some ideas:
  - Just plop your question, verbatim, into Google and see what you get. Include the package you're using and "R". Ex. "how to change legend name in ggplot2 in r". Try typing this in google. You should get a few excellent hits:



stackoverflow.com › questions › ho... ▾ Traduire cette page

How to change legend title in ggplot - Stack Overflow

14 août 2016 — If you were using **a** different type of plot where the code was **ggplot**(data, colour= cond), then you could use +labs(colour= **"Title Name"**) . In ...

12 réponses

Editing **legend** (text) labels in **ggplot**	2 réponses	20 juill. 2016
How can I **change** the **title** of **a ggplot2 legend**?	4 réponses	27 sept. 2016
How to change labels (**legends**) in **ggplot**?	2 réponses	10 août 2020
Changing legend title ggplot	1 réponse	9 janv. 2020

Autres résultats pour stackoverflow.com

cookbook-r.com › Graphs › Legend... ▾ Traduire cette page

Legends (ggplot2) - Cookbook for R

Aller à **Changing** the factor in the data frame — There are two ways of **changing** the **legend title** and labels. The first way is to tell the scale to use have **a** ...

Solution · Changing the order of items ... · Modifying the text of legend ...

- Get comfortable searching **Stack Overflow**, and learning how to apply non-epi solutions to your epi problems.
- If you are using the tidyverse, there are a TON of resources, including Cookbook for R, R for Data Science, etc.

Tips:

1. Start your search statement with the software name (and version as needed)
2. Follow-up with question, function, or error message
    a. In relation to error messages, they are sometimes long and in such cases it may be most useful to take the more generic portions of the returned message
3. Identify and make note of sources/websites that keep popping up that have useful information, such as:
    - https://stackoverflow.com/
    - https://www.rdocumentation.org
    - https://www.tidyverse.org/
    - https://www.dummies.com/programming/r/r-for-dummies-cheat-sheet/
    - http://www.cookbook-r.com/
    - https://www.reconlearn.org/

    Tutorial on troubleshooting and common errors: https://ourcodingclub.github.io/tutorials/troubleshooting/

Knowledge Boost: Silly Mistakes We All Make in R/R Studio

Duration: **4 minutes**

This YouTube video reviews some basic programming challenges encountered when running code in R/RStudio, specifically unbalanced parentheses and quotation marks. “When in doubt, escape [Esc] out”: https://www.youtube.com/watch?v=xQ9SJvuzg0A

Knowledge Boost: Common R Errors

This webpage reviews some errors that are commonly encountered when working with R. Some of the content of this page may be inaccessible to novice or beginner R users. If that is the case we recommend (1) a high level review of the webpage, and (2) filing the webpage away for later use as a resource in your R learning journey when these errors are encountered: https://www.programmingr.com/r-error-messages/

SELF-STUDY EXERCISES

After all these self-learning exercises, it only seems natural that you should test your knowledge at least a little bit! Please complete the following self-reflective and knowledge check questions to ensure you're ready to start the course.

1. In your experience, how is the R programming language different from other statistical software you've used in your work? How is it similar?

2. What is the function for installing packages in R?

3. Using this function, please install the following packages:

   - □ here
   - □ tidyverse
   - □ scales
   - □ padr
   - □ writexl
   - □ fs
   - □ RColorBrewer
   - □ ggrepel
   - □ ggpubr
   - □ zoo
   - □ viridis
   - □ igraph
   - □ tidygraph
   - □ ggraph
   - □ flextable
   - □ viridis
   - □ incidence
   - □ officer
   - □ officedown

4. The R package “tidyverse” is unique, in that it is actually a collection of several packages that you will use frequently throughout this course (and hopefully in your career!) In the space below, please list the different packages included in the “tidyverse”.

“Core Packages”	“Non-core Packages”
1.	1.
2.	2.
3.	3.
4.	4.
5.	5.
6.	6.
7.	7.
8.	8.
	9.
	10.
	11.

5. Did you encounter problems installing R packages? If yes, did you understand what the error message was telling you? If not, did you understand what the error message was telling you after you googled it?

6. List two different ways to look up help documentation for R packages and functions:

    1.
    2.

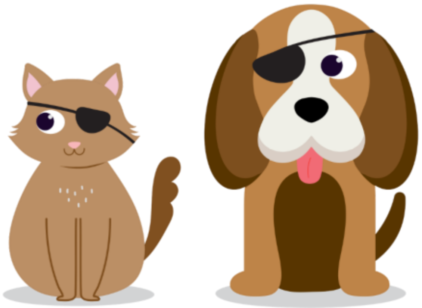
7. How do statistical software handle dates generally? Why is it useful for dates to be handled in this way?

8. What is R Markdown? Can you see any added value of the functionality provided by R Markdown in your day-to-day work? Why or why not?

9. What does the “gg” stand for in the package “ggplot”? Why do you think it’s called this?

10. What is the difference between long and wide format? Provide an example of when you could use each of these formats.

GOING FURTHER



The materials provided below are not in any way required for participation in the Introduction to R for Public Health Investigations course. These links are provided merely as a curated list to help you further your R knowledge based on your own interest level. Caution: using the following material may result in becoming an absolute swashbuckler in R.

There Be Treasure: Going Further

- Applied Epi – Interactive R Tutorials: https://appliedepi.org/tutorial/
- The Epidemiologist R Handbook: https://epirhandbook.com/
- R for Data Science: https://r4ds.had.co.nz/
- R Cheat Sheets: https://rstudio.com/resources/cheatsheets/
- YaRrr! The pirate's guide to R: https://bookdown.org/ndphillips/YaRrr/
- R Shiny Tutorials: https://shiny.rstudio.com/tutorial/
- Comprehensive R Archive Network: https://cran.r-project.org/
- REpidemics Consortium: https://www.repidemicsconsortium.org/projects/
- R4Epis: https://r4epis.netlify.app/
- PHAC R user group (note that this is available to PHAC employees only at this time and you will need to create an account or login with an existing account): https://message.gccollab.ca/channel/phac-r-user
- PopDataBC – Intro to R Studio for SAS Users: https://www.popdata.bc.ca/etu/online_courses/STAN104
- PopDataBC - Data Management and Cleaning for Analysis with R: https://www.popdata.bc.ca/etu/online_courses/STAN106
- A comprehensive introduction to handling date and time in R: https://blog.rsquaredacademy.com/handling-date-and-time-in-r/
- Regular tips, tricks, and demonstrations of new packages: https://www.r-bloggers.com/
- Stack overflow: https://stackoverflow.com/
- Coursera - Introduction to Statistics and Data Analysis in R: https://www.coursera.org/specializations/statistical-analysis-r-public-health?recoOrder=1&utm_medium=email&utm_source=recommendations&utm_campaign=btfwMI53EeuiLbV1NbTVAw

(Note: if experiencing issues accessing a link by clicking on it, please try to troubleshoot via copy-paste into your browser.)

APPENDIX 1: PRE-COURSE CHECKLIST

Introduction to R for Public Health Investigations

Actions to be completed prior to joining the virtual classroom on Day 1	Completed
Items to complete 1 to 3 weeks before the course	
Complete the knowledge check and associated pre-learning in the Introduction to R for Public Health Investigations Self-Study Module	□
Complete the pre-course self-assessment appended to the Introduction to R for Public Health Investigations Self-Study Module	□
Items to complete 3 to 6 days before the course	
Advise the TDU if you are unable to access the Participant Guide located on the file share platform for this course. Note that the link will follow in an email and course content will be provided one week prior to the course.	□
Install R and R Studio if joining from non-PHAC computer (please refer to the associated tutorial in the Introduction to R For Public Health Investigations Pre-Course Self-Study Module as needed).	□
Test virtual classroom (Zoom) audio and video on the computer you will be using for the duration of the course: https://support.zoom.us/hc/en-us/articles/115002262083-Joining-a-test-meeting	□
Advise supervisor and coworkers that you will be unavailable for the duration of the course.	□
OPTIONAL: Print course materials for note taking as needed.	□
Items to complete 1 to 3 days prior the course	
Ensure that you have reviewed pre-learning for Day 1 (refer to the Introduction to R for Public Health Investigations Participant Guide).	□
Download materials for the practical exercise on Day 1 and advise the TDU of any issues.	□
Advise your supervisor and coworkers that you will be unavailable for the duration of each virtual classroom session (incl. ~15 minutes before and after).	□
Locate the Zoom meeting details for the virtual classroom and office hours, and ensure that they will be easy to find during the course.	□
Advise the TDU if you are unable to access the Slack workspace for this training. Note that the link to access this Slack workspace will follow in an email.	□
Steps to complete 15 minutes prior to joining each virtual session	
Turn off all telephone notifications, and set wireless devices to silent.	□
Prepare your computer by closing all non-required applications (incl. email and instant messaging).	□
Access the virtual classroom 15 minutes prior to each session to resolve potential issues and get ready on time.	□

Please contact the Training and Development Unit at ceptraining-formationcmu@phac-aspc.gc.ca for assistance.

APPENDIX 2: PRE- AND POST-COURSE SELF-ASSESSMENT

This self-assessment is intended to be a self-reflective exercise for you to complete ahead of and following the course. The same reflection question is provided for you to complete just prior to joining the virtual classroom, and again after you've completed the course and associated exercises. This activity is optional. However, we recommend that field epidemiologists in the Canadian Field Epidemiology Program complete this activity and use it to inform discussions with their program in relation to remaining individual gaps and needs as they relate to technical skill proficiency.

PRE-COURSE SELF ASSESSMENT

Before joining the Introduction to R for Public Health Investigations course, how confident do you feel developing R code by yourself for the following tasks (1= not confident; 10 = very confident)?

Skill	Not confident									Very confident
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Setting a working directory										
Reshaping data										
Creating new variables										
Identifying missing data										
Descriptive epidemiology										
Creating an epidemic curve										
Importing a dataset										
Merging and appending data										
Automating a report										

POST-COURSE SELF ASSESSMENT

After completing the Introduction to R for Public Health Investigations course, how confident do you feel developing R code by yourself for the following tasks (1= not confident; 10 = very confident)?

Skill	Not confident									Very confident
	1	2	3	4	5	6	7	8	9	10
Setting a working directory										
Reshaping data										
Creating new variables										
Identifying missing data										
Descriptive epidemiology										
Creating an epidemic curve										
Importing a dataset										
Merging and appending data										
Automating a report										