PH 251D Applied Epidemiology Using R, Fall 2019

Tomás J. Aragón, MD, DrPH[1,2,3,4,*]

[1] Health Officer, City & County of San Francisco
[2] Director, Population Health Division, SFDPH
[3] Clinical faculty, UC Berkeley School of Public Health, Epidemiology
[4] Affiliate faculty, UCSF Department of Epidemiology & Biostatistics
* Email: aragon@berkeley.edu, Cell: 415-515-5734
Web: https://taragonmd.github.io/

Updated: August 31, 2019

1 Basic Information

- Course: Public Health 251D
- Course title: Applied Epidemiology Using R
- Day, time, and location: Fridays, 4pm–6pm, Berkeley Way West 1203
- Course Control Number: 29489 (auditors welcome if there is seating room)
- Units/credits: 2 units
- Office hours: TBD
- Course syllabus: https://github.com/taragonmd/teaching/ (ph251d folder)
- Instructors:

Name	Affiliation	Email	Phone
Tomás J. Aragón	SF Dept of Public Health	aragon@berkeley.edu	415.515.5734
Michael Samuel	Ca. Dept of Public Health	michael.samuel@sfdph.org	925.285.2926
Wayne Enanoria	SF Dept of Public Health	wayne.enanoria@sfdph.org	415.437.6356

2 Course Description

2.1 Prerequisites

- Prior courses: introductory courses in epidemiology and Biostatistics
- Prior experience: basic epidemiologic programming (e.g., Stata, SAS)
- Motivation: relentless commitment to invest time coding in R

2.2 Overview of course

This is an intensive one-semester introduction to the R programming language for applied epidemiology from a population health data science perspective. **Population health** is "a systems framework for studying and improving the health of populations through collective action and learning" [1]. Populations include clinical and community populations (neighborhoods, social networks). **Population health data science** (PHDS) is "the art and science of transforming data into information and actionable knowledge that informs, influences or optimizes decisions[1] to protect and improve the health and well-being of populations."

PHDS rests on the foundation of **population health thinking** (PHT) [2]—a problem-solving, inferential framework that uses Bayesian networks for

1. probabilistic reasoning,
2. causal inference, and
3. decision quality.

[1] A *decision* is an irrevocable commitment of time or resources towards achieving an objective. In public health, and in life, decision making is our single, most important activity. Every decision has an opportunity cost—the loss benefit of the better option not chosen or not considered. "Decisions determine destiny" (Frederick Speakman).

PHT supports causal, evidential, and inter-causal reasoning. BNs, and it variants (directed acyclic graphs, decision networks), are used for causal thinking, study design, counterfactuals, deconfounding, decision analysis, and more. R is the ideal programming language to learn PHT/PHDS at an introductory level which supports public health analysts, practitioners, and researchers to bring value, insights, and causal rigor to the trans-disciplinary table of data science (with computer scientists, mathematicians, statisticians, subject matter experts, etc.).

The core of population health data science is the timely analysis and synthesis of data using programming and computing power. Fortunately for us we have R! R is a freely available, multi-platform (Linux, Mac OS, Windows, etc.), versatile, and powerful program for statistical computing and graphics (http://www.r-project.org). This course will focus on core basics of organizing, managing, and manipulating population health data; introduction to R programming; and basic R graphics, all applied to PHT/PHDS. Students will complete and present a project in their field of interest.

PHDS has five analytic domains (Figure 1 and Table 2):

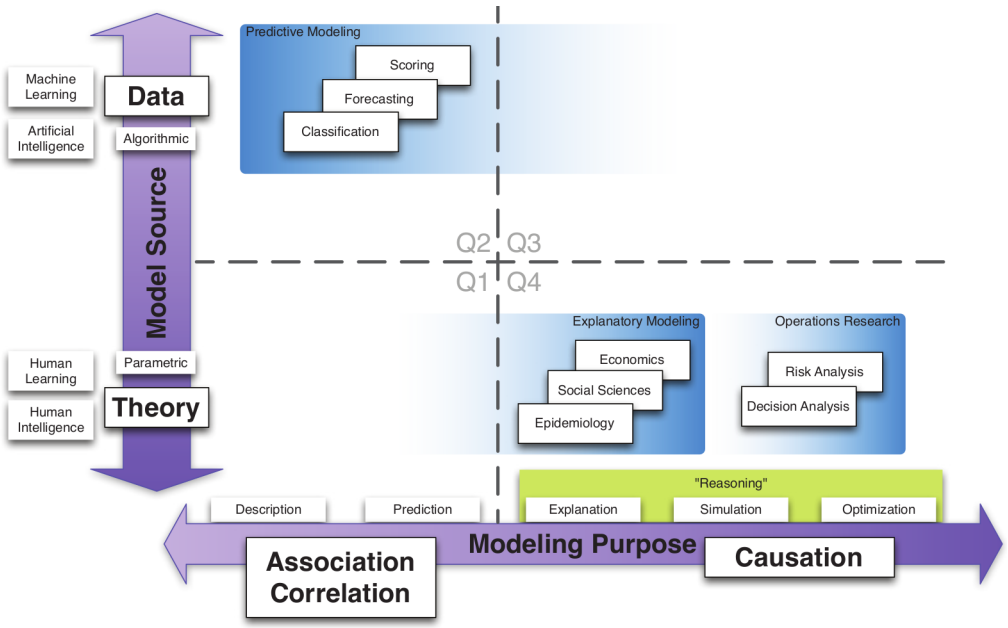

Figure 1. The population health data science landscape (source: www.bayesia.com)

Table 2. Population health data science: Levels of analysis

Level	Analytic domain	Description
1	Description	surveillance and early detection of events prevalence and incidence of risks and outcomes
2	Prediction	early prediction and targeting of interventions
3	Explanation (causal inference)	discovery of new causal effects and pathways estimation of intervention effectiveness
4	Simulation	modeling for epidemiologic or decision insights
5	Optimization[1]	informing or optimizing decisions, processes, or results

2.3 Intended audience

This course is intended for epidemiologists, medical epidemiologists, data analysts, and demographers that want an introduction to the R language for population health data science applications.

2.4 Methods of instruction

Lecture and computer demonstration. You are welcome to bring your laptop with R pre-installed.

3 Schedule

Day	Dates	Wk	Content or milestone	Speaker
Wed	Aug 21		FALL SEMESTER BEGINS	
Wed	Aug 28		University instruction Begins	
Fri	Aug 30	1	**Chap 1:** Getting started with R (1st day)	
Mon	Sep 03		*Academic and Administrative Holiday*	
Fri	Sep 06	2	**Chap 2:** Working with R data objects 1	TJA
Fri	Sep 13	3	**Chap 3:** Working with R data objects 2	TJA
Fri	Sep 20	4	**Chap 4:** Managing population health data 3	TJA
Fri	Sep 27	5	**Chap 4:** Managing population health data 4	TJA
Fri	Oct 04	6	**Chap 5:** R programming 1	TJA
Fri	Oct 11	7	**Chap 5:** R programming 2	TJA
Fri	Oct 18	8	Graphing population health data 1	MCS
Fri	Oct 25	9	PH Data Science Analytics 1	TJA
Fri	Nov 01	10	Graphing population health data 2	MCS
Fri	Nov 08	11	PH Data Science Analytics 2	TJA
Fri	Nov 15	12	PH Data Science Analytics 3	TJA
Fri	Nov 22	13	PH Data Science Analytics 4	TJA
Thu-Fri	Nov 28–29		*Academic and Administrative Holiday*	
Fri	Dec 06	14	Student project presentations	TJA
Mon-Fri	Dec 09-13		Reading/Review/Recitation Week	
Mon-Fri	Dec 16–20		Final Examinations week	

4 Learning outcomes/goals/objectives

1. Use R as a scientific calculator and a functional spreadsheet;
2. Enter, manage, and manipulate population health data in R;
3. Conduct basic population health analyses in R;
4. Graphically display population health data;
5. Write basic R functions (programs);
6. Learn basic literate programming using R markdown; and
7. Understand how to construct simple Bayesian networks for probabilistic reasoning, causal inference, and decision analysis.

5 Materials

5.1 Books

1. Tomás J Aragón (2019). *Population Health Data Science with R—Transforming data into actionable knowledge*. Available from: https://bookdown.org/taragonmd/phds/ (**REQUIRED**).
2. R Foundation. *An Introduction to R*, Freely available at http://cran.r-project.org/doc/manuals/R-intro.pdf (also comes with default installation) (**REQUIRED**)
3. Judea Pearl & Dana Mackenzie (2018). *The Book of Why: The new science of cause and effect*. Basic Books; 1 ed. (highly recommended)
4. Judea Pearl, Madelyn Glymour, Nicholas P. Jewell (2016). *Causal Inference in Statistics: A Primer*. Wiley (recommended)

5. Marco Scutari (2014). *Bayesian Networks: With Examples in R* Chapman and Hall/CRC (recommended)

5.2 Data sets

You can access data for this course from here:

- https://github.com/taragonmd/data (download data sets used in book chapters)
- http://methods.sagepub.com/datasets

5.3 Software

1. Install R on your computer (visit http://cran.cnr.berkeley.edu/)
2. Install RStudio on your computer (visit http://www.rstudio.com)
3. Set up RPubs account at Rpubs.com (visit http://rpubs.com/)
4. Creating LaTeX documents (optional): use LaTeX to create high quality typeset documents.

RStudio also generates LaTeX documents, but it must be installed on your computer. For MS Window, install MikTeX. For Mac OS, install MacTeX. For Linux, install TeX Live. Alternatively, install the `tinytex` package in R. `tinytex` is a cross-platform helper functions to install and maintain TeX Live, and compile LaTeX documents.

5.4 Websites

1. bCourses: https://bcourses.berkeley.edu/ (UC Berkeley Learning Management System)
2. PHDS: https://taragonmd.github.io/project/teaching/

6 Requirements: exams, quizzes, assignments

6.1 Assignments

1. Complete Homework assignments.
2. Complete 2 projects (see below)

6.2 Homework

1. Problems will be assigned from the end of each chapter.
2. Homework will be due 1 week after I finish covering chapter in class. Check bCourses.
3. Submit complete homework as described in bCourses.

6.3 Project 1

TBD

6.4 Project 2

TBD

7 Policies: grading procedures, attendance, participation, etc

7.1 Grading

For registered UC Berkeley/Extension students: Units: 2; Grading: Letter or S/U

Grading will be based on

- 30% Homework,
- 30% Student Project 1, and
- 40% Student Project 2.

Satisfactory (S) or Passed (P) is at a minimum level of B.

Table 4. Letter grading scale

Percent	Grade	Percent	Grade
100–94	A	75–73	C
93–90	A-	72–70	C-
89–86	B+	69–66	D+
85–83	B	65–63	D
82–80	B-	62–60	D-
79–76	C+	< 60	F

7.2 Attendance

Attendance is very important to learn the material. Please provide feedback on how I can improve my instruction of the material.

7.3 R resources

- Official R manuals: http://cran.r-project.org/manuals.html
- Contributed R tutorials: http://cran.r-project.org/other-docs.html
- UCB Intelligent research design for data intensive social science: http://dlab.berkeley.edu/
- Resources to help you learn and use R: https://stats.idre.ucla.edu/r/
- R User Groups: http://r-users-group.meetup.com/
- Bay Area useR Group (R Programming Language): https://www.meetup.com/R-Users/
- The East Bay R Language Beginners Group: https://www.meetup.com/r-enthusiasts/

8 Resources: tips for success, glossaries, links, academic support services, etc

8.1 Academic Calendar and Student Accommodations - Campus Policies and Guidelines

https://teaching.berkeley.edu/academic-calendar-and-student-accommodations-campus-policies-and-guidelines

8.2 Student resources

Available here: https://sph.berkeley.edu/current-students/student-resources

8.3 Policies for all Courses offered at SPH

Available here: https://berkeley.app.box.com/s/knh3rbk9ikgvmca4ymy93msgj9bkebq5

9 Safety and Emergency Preparedness

In the event that we need to evacuate the building we will gather and meet at the TBD side of the building.

10 UC Berkeley Honor Code (link is external)

The Honor Code is available here: https://teaching.berkeley.edu/berkeley-honor-code

References

1. Aragón TJ, Colfax G. We will be the best at getting better! A playbook for population health improvement. UC Berkeley eScholarship [Internet]. 2019; Available from: https://escholarship.org/uc/item/9xg5t30s

2. Aragón TJ. Population health thinking using Bayesian networks. University of California, eScholarship [Internet]. 2018; Available from: https://escholarship.org/uc/item/8000r5m5