JOHNS HOPKINS
BLOOMBERG SCHOOL
of PUBLIC HEALTH

Introduction

Lecture 1

Outline

- Why are you here
- What material will be covered
- Who we are
- What to expect

Why are you here?

- For most of you – you are required to be!

Why are you here?

- What have you done to date…
- …and where we are going in this class.

Causal Inference

- Measuring outcomes
  - Assessing and calculating burden
- Understanding exposures
  - How do we define them?
  - How we measure them?
- Epidemiologic Inquiry
  - How (if) do the exposure(s) relate to the outcome?
  - What are design methods and analytic methods we can apply to answer this question?

This Class

- Measuring outcomes
  - Assessing and calculating burden
- Understanding exposures
  - How do we define them?
  - How we measure them?
- Epidemiologic Inquiry
  - How (if) do the exposure(s) relate to the outcome?
  - What are design methods and analytic methods we can apply to answer this question?

Vision: Describe population health to optimize opportunities for intervention

- How do we utilize data from national surveys to draw inferences about risk factors affecting the health of various subpopulations in the US?
- How do we quantify potential improvements in core health and well-being measures if certain health risks could be reduced?
- How do we frame our inquiry so that we understand where social and cultural factors fit in, and although no one study needs to “solve all problems”, how do we look at the big picture so that we narrow our scope appropriately?

Vision: Describe population health to optimize opportunities for intervention

- How do we utilize data from national surveys to draw inferences about risk factors affecting the health of various subpopulations in the US?
- How do we quantify potential improvements in core health and well-being measures if certain health risks could be reduced?
- How do we frame our inquiry so that we understand where social and cultural factors fit in, and although no one study needs to "solve all problems", how do we look at the big picture so that we narrow our scope appropriately?
- Who uses these methods/skills?
  - Both applied and research epidemiologists

Module 1

Learning Objectives: Weighted Survey Analysis: Analytic techniques for the incorporation of weights in the analysis of survey data to make inferences about the target population.

PLOS ONE

RESEARCH ARTICLE

Implementing machine learning methods with complex survey data: Lessons learned on the impacts of accounting sampling weights in gradient boosting

Nathaniel MacNell[1], Lydia Feinstein[1,2], Jesse Wilkerson[1], Päivi M. Salo[3], Samantha A. Molsberry[1], Michael B. Fessler[3], Peter S. Thorne[4], Alison A. Motsinger-Reif[3], Darryl C. Zeldin[3]*

1 Social & Scientific Systems, a DLH Holdings Company, Durham, North Carolina, United States of America, **2** Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina, Chapel Hill, North Carolina, United States of America, **3** Division of Intramural Research, National Institute of Environmental Health Sciences, National Institutes of Health, Durham, North Carolina, United States of America, **4** Department of Occupational and Environmental Health, University of Iowa, College of Public Health, Iowa City, Iowa, United States of America

* zeldin@niehs.nih.gov

OPEN ACCESS

Citation: MacNell N, Feinstein L, Wilkerson J, Salo PM, Molsberry SA, Fessler MB, et al. (2023) Implementing machine learning methods with complex survey data: Lessons learned on the impacts of accounting sampling weights in gradient boosting. PLoS ONE 18(1): e0280387. https://doi.org/10.1371/journal.pone.0280387

Editor: Kathiravan Srinivasan, Vellore Institute of Technology: VIT University, INDIA

Received: September 28, 2022

Accepted: December 28, 2022

Published: January 13, 2023

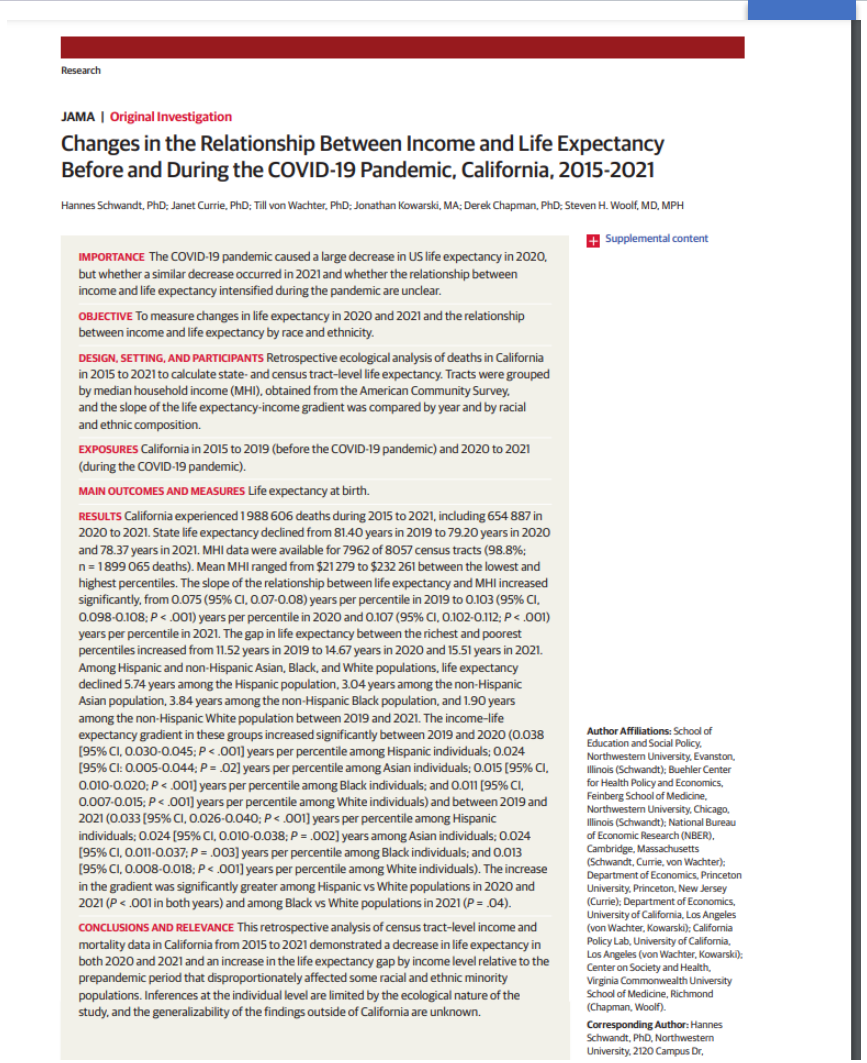
Abstract

Despite the prominent use of complex survey data and the growing popularity of machine learning methods in epidemiologic research, few machine learning software implementations offer options for handling complex samples. A major challenge impeding the broader incorporation of machine learning into epidemiologic research is incomplete guidance for analyzing complex survey data, including the importance of sampling weights for valid prediction in target populations. Using data from 15, 820 participants in the 1988–1994 National Health and Nutrition Examination Survey cohort, we determined whether ignoring weights in gradient boosting models of all-cause mortality affected prediction, as measured by the F1 score and corresponding 95% confidence intervals. In simulations, we additionally assessed the impact of sample size, weight variability, predictor strength, and model dimensionality. In the National Health and Nutrition Examination Survey data, unweighted model performance was inflated compared to the weighted model (F1 score 81.9% [95% confidence interval: 81.2%, 82.7%] vs 77.4% [95% confidence interval: 76.1%, 78.6%]). However, the

Module 2

Learning Objectives:

- Calculating Life Expectancy: Calculate a single-decrement life table, and how to create them using compilation commands as well as publicly available Excel-based tools.



Research

JAMA | Original Investigation

Changes in the Relationship Between Income and Life Expectancy Before and During the COVID-19 Pandemic, California, 2015-2021

Hannes Schwandt, PhD; Janet Currie, PhD; Till von Wachter, PhD; Jonathan Kowarski, MA; Derek Chapman, PhD; Steven H. Woolf, MD, MPH

IMPORTANCE The COVID-19 pandemic caused a large decrease in US life expectancy in 2020, but whether a similar decrease occurred in 2021 and whether the relationship between income and life expectancy intensified during the pandemic are unclear.

OBJECTIVE To measure changes in life expectancy in 2020 and 2021 and the relationship between income and life expectancy by race and ethnicity.

DESIGN, SETTING, AND PARTICIPANTS Retrospective ecological analysis of deaths in California in 2015 to 2021 to calculate state- and census tract–level life expectancy. Tracts were grouped by median household income (MHI), obtained from the American Community Survey, and the slope of the life expectancy-income gradient was compared by year and by racial and ethnic composition.

EXPOSURES California in 2015 to 2019 (before the COVID-19 pandemic) and 2020 to 2021 (during the COVID-19 pandemic).

MAIN OUTCOMES AND MEASURES Life expectancy at birth.

RESULTS California experienced 1 988 606 deaths during 2015 to 2021, including 654 887 in 2020 to 2021. State life expectancy declined from 81.40 years in 2019 to 79.20 years in 2020 and 78.37 years in 2021. MHI data were available for 7962 of 8057 census tracts (98.8%; n = 1 899 065 deaths). Mean MHI ranged from $21 279 to $232 261 between the lowest and highest percentiles. The slope of the relationship between life expectancy and MHI increased significantly, from 0.075 (95% CI, 0.07-0.08) years per percentile in 2019 to 0.103 (95% CI, 0.098-0.108; $P < .001$) years per percentile in 2020 and 0.107 (95% CI, 0.102-0.112; $P < .001$) years per percentile in 2021. The gap in life expectancy between the richest and poorest percentiles increased from 11.52 years in 2019 to 14.67 years in 2020 and 15.51 years in 2021. Among Hispanic and non-Hispanic Asian, Black, and White populations, life expectancy declined 5.74 years among the Hispanic population, 3.04 years among the non-Hispanic Asian population, 3.84 years among the non-Hispanic Black population, and 1.90 years among the non-Hispanic White population between 2019 and 2021. The income–life expectancy gradient in these groups increased significantly between 2019 and 2020 (0.038 [95% CI, 0.030-0.045; $P < .001$] years per percentile among Hispanic individuals; 0.024 [95% CI: 0.005-0.044; $P = .02$] years per percentile among Asian individuals; 0.015 [95% CI, 0.010-0.020; $P < .001$] years per percentile among Black individuals; and 0.011 [95% CI, 0.007-0.015; $P < .001$] years per percentile among White individuals) and between 2019 and 2021 (0.033 [95% CI, 0.026-0.040; $P < .001$] years per percentile among Hispanic individuals; 0.024 [95% CI, 0.010-0.038; $P = .002$] years among Asian individuals; 0.024 [95% CI, 0.011-0.037; $P = .003$] years per percentile among Black individuals; and 0.013 [95% CI, 0.008-0.018; $P < .001$] years per percentile among White individuals). The increase in the gradient was significantly greater among Hispanic vs White populations in 2020 and 2021 ($P < .001$ in both years) and among Black vs White populations in 2021 ($P = .04$).

CONCLUSIONS AND RELEVANCE This retrospective analysis of census tract–level income and mortality data in California from 2015 to 2021 demonstrated a decrease in life expectancy in both 2020 and 2021 and an increase in the life expectancy gap by income level relative to the prepandemic period that disproportionately affected some racial and ethnic minority populations. Inferences at the individual level are limited by the ecological nature of the study, and the generalizability of the findings outside of California are unknown.

Supplemental content

Author Affiliations: School of Education and Social Policy, Northwestern University, Evanston, Illinois (Schwandt); Buehler Center for Health Policy and Economics, Feinberg School of Medicine, Northwestern University, Chicago, Illinois (Schwandt); National Bureau of Economic Research (NBER), Cambridge, Massachusetts (Schwandt, Currie, von Wachter); Department of Economics, Princeton University, Princeton, New Jersey (Currie); Department of Economics, University of California, Los Angeles (von Wachter, Kowarski); California Policy Lab, University of California, Los Angeles (von Wachter, Kowarski); Center on Society and Health, Virginia Commonwealth University School of Medicine, Richmond (Chapman, Woolf).

Corresponding Author: Hannes Schwandt, PhD, Northwestern University, 2120 Campus Dr,

Module 3

Learning Objectives:

- Estimate Preventable Deaths: Calculate summary measures of mortality, and econometric techniques for estimating lives saved resulting from modification of risk factors.

International Journal of Epidemiology, 2019, 1367–1376
doi: 10.1093/ije/dyy254
Advance Access Publication Date: 9 January 2019
Education Corner

Education Corner

Reflection on modern methods: years of life lost due to premature mortality—a versatile and comprehensive measure for monitoring non-communicable disease mortality

Ramon Martinez,[1]* Patricia Soliz,[2] Roberta Caixeta[1] and Pedro Ordunez[1]

[1]Department of Non-Communicable Diseases and Mental Health and [2]Department of Evidence and Intelligence for Action in Health, Pan American Health Organization, Washington, DC, USA

*Corresponding author. Department of Non-Communicable Diseases and Mental Health, Pan American Health Organization, 525 23 rd St NW, Washington, DC 22037, USA. E-mail: martiner@paho.org

Editorial decision 11 October 2018; Accepted 3 December 2018

Abstract

The analysis of causes impacting on premature mortality is an essential function of public health surveillance. Diverse methods have been used for accurately assessing and reporting the level and trends of premature mortality; however, many have important limitations, particularly in capturing actual early deaths. We argue that the framework of years of life lost (YLL), as conceptualized in disability-adjusted life-years (DALYs), is a robust and comprehensive measure of premature mortality. Global Burden of Disease study is systematically providing estimates of YLL; however, it is not widely adopted at country level, among other reasons because its conceptual and methodological bases seem to be not sufficiently known and understood. In this paper, we provide the concepts and the methodology of the YLL framework, including the selection of the loss of function that defines the time lost due to premature deaths, and detailed methods for calculating YLL metrics. We also illustrate how to use YLL to quantify the level and trends of premature non-communicable disease (NCD) mortality in the Americas. The tutorial style of the illustrative example is intended to educate the public health community and stimulate the use of YLL in disease prevention and control programmes at different levels.

Key words: Mortality, premature, epidemiological method, non-communicable diseases, public health surveillance

Module 4

Learning Objectives:

- Application of Conceptual Frameworks in Epidemiology: Understand the strengths and weaknesses of various commonly used frameworks, and an introduction to translational epidemiology.

American Journal of Epidemiology

Vol. 188, No. 12
DOI: 10.1093/aje/kwz085
Advance Access publication: March 30, 2019

Commentary

From Epidemiologic Knowledge to Improved Health: A Vision for Translational Epidemiology

Michael Windle, Hojoon D. Lee, Sarah T. Cherng, Catherine R. Lesko, Colleen Hanrahan, John W. Jackson, Mara McAdams-DeMarco, Stephan Ehrhardt, Stefan D. Baral, Gypsyamber D'Souza, and David W. Dowdy*

* Correspondence to Dr. David W. Dowdy, Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, 615 North Wolfe Street, Suite E6531, Baltimore, MD 21205 (e-mail: ddowdy1@jhmi.edu).

Initially submitted September 5, 2018; accepted for publication March 21, 2019.

Epidemiology should aim to improve population health; however, no consensus exists regarding the activities and skills that should be prioritized to achieve this goal. We performed a scoping review of articles addressing the translation of epidemiologic knowledge into improved population health outcomes. We identified 5 themes in the translational epidemiology literature: foundations of epidemiologic thinking, evidence-based public health or medicine, epidemiologic education, implementation science, and community-engaged research (including literature on community-based participatory research). We then identified 5 priority areas for advancing translational epidemiology: 1) scientific engagement with public health; 2) public health communication; 3) epidemiologic education; 4) epidemiology and implementation; and 5) community involvement. Using these priority areas as a starting point, we developed a conceptual framework of translational epidemiology that emphasizes interconnectedness and feedback among epidemiology, foundational science, and public health stakeholders. We also identified 2–5 representative principles in each priority area that could serve as the basis for advancing a vision of translational epidemiology. We believe an emphasis on translational epidemiology can help the broader field to increase the efficiency of translating epidemiologic knowledge into improved health outcomes and to achieve its goal of improving population health.

education; evidence-based medicine; translational medical research

In many recent commentaries in the epidemiology literature, the focus has been on the role of epidemiologists in not only identifying causes of disease but also in improving population health (1–4). Emerging methods in epidemiology, including agent-based modeling (5) and target trial emulation (6), are likewise increasingly focused on answering questions related to the improvement of health, as well as identifying etiologies of disease. Very few epidemiologists would take issue with the assertion that epidemiology should provide evidence that ultimately aims to advance the health of populations. Little consensus exists, however, regarding how best to translate knowledge from epidemiologic studies into better population health.

Many proposals have advocated for specific improvements to analytic methods (7), education (8), and public health practice (9); however, we still lack a coherent vision for answering the question: If epidemiology aims to improve population health, what activities and skills should epidemiologists prioritize? A holistic discussion of this question is particularly timely, because epidemiologists increasingly face struggles to sustain funding streams, compete for high-quality students, maintain relevance in the public eye, bridge gaps between academic and public health efforts, and improve scientific rigor.

Here, we endeavor to link epidemiologic knowledge and population health outcomes using the lens of translational epidemiology, defined by Szklo (10) as the "effective transfer of new knowledge from epidemiologic studies into the planning of population-wide and individual-level disease control programs and policies." To help crystallize a vision of translational epidemiology, we conducted a scoping review of existing literature addressing this topic. On the basis of those results, we propose a conceptual framework of translational epidemiology and synthesize recommendations from the reviewed literature toward developing a more translational modern epidemiology.

Who we are

- Lead Faculty: Aruna Chandran
- MD Degree: Johns Hopkins School of Medicine
- Residency in Pediatrics: Northwestern's Children's Memorial Hospital
- Fellowship in Public Health and Preventive Medicine: Johns Hopkins Bloomberg School of Public Health
- Clinical practice: Pediatric emergency department at urgent care at Bayview Medical Center

- International Health: Infectious Diseases and Injury/Violence Prevention
  - Establishing surveillance systems
  - Program evaluation
  - Clinical trials
- Baltimore City Health Department: Chief of Epidemiologic Services
- Epidemiology
  - Social/Structural Determinants of Health
  - HIV prevention/care
  - Child health outcomes

Awesome TAs

- Nandita Somayaji
- Darpa Anireddy
- Both 2nd year Master's degree candidates in the Department of Epidemiology

Course Structure: Synchronous Online Instruction

- Feedback: Challenges of making it to in-person classes at 8:30
- There will be a synchronous didactic lecture followed by a synchronous TA session nearly every day.
  - Check course website for schedule and links
- Our goal is for you to appreciate/understand the material/concepts presented.
  - We recognize that there will be challenges
- We will post recordings of all lectures after each class.

Course Structure: Lectures

- Core Lectures: Presentation of core course material in a synchronous fashion
  - All core lectures (given by Dr. Aruna Chandran) will be presented this way
  - Will be recorded and then posted
  - Post questions in the “Lecture Questions” section of the Discussion Forum
- TA Sessions: Synchronous online.
  - For TA Sessions through Wednesday April 24th, attend the one corresponding to the statistical program you are using
- Readings:
  - Optional readings posted in the Online library

Assignments: Logistics

- Four homework assignments (1 for each module)
  - Lecture slides will describe analytic techniques
- Each assignment MUST be completed individually
  - Can discuss assignments/answers as well as collaborate (for software issues)
- Designed as “quizzes” in courseplus, which you may access multiple times

Assignments

- Due Dates: (all assignments due on Wednesdays by 11:59pm)
  - Assignment 1: April 10
  - Assignment 2: April 24
  - Assignment 3: May 8
  - Assignment 4: May 15
- Late assignments: Please do your best to turn things in on time, or be in touch with us if there is a problem

Assignments: Software

- Assignments: You are welcome to work in either Stata or R
  - You have faculty/TA support for both
    - Darpa codes in R, Nandita codes in Stata
- "Coding supplements" are provided for Modules 1, 2 and 3 (not needed for 4)
  - Give detailed code in Stata and R (you have to insert variable names, *etc.*)
  - Give detailed instructions for mapping
- You are NOT being supplied with do-files
  - You do need to write your analytic commands/code

Assignments: TA Help

- TAs will be running synchronous TA sessions from 9:30-10:20 nearly every class session (except first and last days)
  - No pre-designated TA “office hours”
- Please post questions in the “Assignments” portion of the Discussion Forum
  - Monitored by the TAs
- Contact TAs by email to set up times to discuss additional questions/issues

Can I use this class to “Learn R”?

- Short answers:
  - I have never opened R before: Probably not
  - I have some basic familiarity with R: Probably yes
- You are NOT being provided with basic code or instructions to read-in a dataset or do simple operations
- You are expected to fill in variable names and other relevant details, which you may not know how to do if you are not familiar with R

Background in Epidemiology

- The class is targeted towards individuals who have completed the Epidemiologic Methods (751-753) OR the Epidemiologic Inference (721-722) sequence.
- We may refer to material covered in the Methods series, but anything you need to know will be reinforced here.
- If you need information (or a refresher) on fundamental epidemiologic concepts beyond what is covered in the lecture, please reach out to us.

Graduation

- Is anyone graduating this May?
- For graduating students: grades are due in the system by Friday, May 17th
  - Therefore, due dates for assignments are the same if you are graduating or not
  - PLEASE email me if you are graduating so we keep you in mind for grading