

# **PUBH 7430**

## **Lecture 1**

Erika Helgeson

Division of Biostatistics  
University of Minnesota School of Public Health

# PUBH 7430: Statistical Methods for Correlated Data

# First Things First

- **Welcome Back!**
- Did you get my email?
- Can you access the course on Canvas?

# Flexible Learning Options

## Flexibility

- Attendance is not mandatory. Lectures will be recorded.
- Zoom and in person office hours
- Everything is on Canvas
- Online quizzes and midterm
- Virtual end of semester presentation
- Additional flexible options as needed

## Classroom Safety

- Stable seating

# Course objectives

- Identify situations where correlated data may arise.
- Describe and summarize correlation in a dataset both graphically and numerically.
- Apply appropriate statistical estimation techniques to answer scientific questions using correlated data.
- Understand both the strengths and limitations of these techniques.
- Work as a team to answer a scientific question.
- Communicate written analysis results in a manner appropriate for a scientific publication.
- Clearly present a scientific problem, analysis, and results using a presentation.

# Course Overview

- Introduction
- Summarizing and visualizing correlated data
- The generalized linear model
- Generalized estimating equations
- Random effects models
- Special topics (time permitting ...)

# Syllabus review

- Pre-requisites
  - Statistics: regression (multiple linear regression and logistic regression), random variables, statistical inference. (PUBH 6451, PUBH 7405, or STAT 5302)
  - Some familiarity with linear algebra
  - Working knowledge of R or SAS
- No textbook
- Class website (**Canvas**)
- Instructor office hours: 11-12 Tuesdays NHH 2-101

**Jiuzhou Wang**  
wang9062@umn.edu



**Damon Leach**  
leach090@umn.edu





# Instructor's personal assistant



# Components of course

- Recorded lectures
- Eight quizzes
- Five graded homework assignments
- Online midterm (Tuesday, Nov. 23)
- Group project
- No in-class final
- See *Syllabus* section on Canvas for more details

# Group Project

Components:

- Survey (used to assign groups)
- Project proposal
- Statistical analysis plan
- Results
- “Show and Tell” virtual presentation
- Written final report
- Peer evaluation

# Key Dates

- Quizzes due roughly every Monday
- Homework assignments roughly every two weeks (Thursdays)
- Project “check-ins” throughout the course
- Midterm online Tuesday, Nov. 23. More information forthcoming
- See *Syllabus* section on Canvas for more details

# Course grade

Grade distribution:

- 35% homework (each assignment contributes equally)
- 20% midterm
- 15% Canvas quizzes (each quiz contributes equally)
- 30% group project

Extra credit for completion of “check-in” quizzes on Canvas

# Getting questions answered

- During class!
- Office hours
- Questions applicable to classmates: Canvas discussion board
- Questions not applicable to classmates: Email me and the TAs

# Homework submission

- Can work together, but must independently write assignments, including any code, in your own words
- Submit assignments on Canvas
- No late assignments will be accepted
- Regrading policy (see syllabus)

# Syllabus Questions?





Let's get started!

# What is correlation?

(Informally:) The degree to which outcomes  
“move together” or are informative about  
each other

# How does correlation arise?

When outcomes share common features or characteristics which we can't model (or simply don't want to)

# Why should we care about correlation?

Not accounting for correlation when it is present can seriously compromise most statistical analyses

# Example 1 of study with correlated data

JAMA Pediatrics | [Original Investigation](#)

## Association of Screen Time and Depression in Adolescence

Elroy Boers, PhD; Mohammad H. Afzali, PhD; Nicola Newton, PhD; Patricia Conrod, PhD

# Questions:

- ① What were the scientific goals/hypotheses of the study?
- ② **What sources of correlation are present in the data?**
- ③ Are you familiar with the methods?
- ④ Carefully read the text describing Table 1 and Table 2. Does the text clearly present the results of their model?
- ⑤ Do the authors include a measure of uncertainty in the text when presenting the results of their analyses?
- ⑥ Look at the tables. Are the captions descriptive? Can you interpret the results?
- ⑦ Are there any other figures or tables you would want to see?
- ⑧ Do you notice something missing from the results?
- ⑨ Can you think of any additional limitations to the study?

# Example 2 of study with correlated data



## The NEW ENGLAND JOURNAL of MEDICINE

[HOME](#)[ARTICLES & MULTIMEDIA ▾](#)[ISSUES ▾](#)[SPECIALTIES & TOPICS ▾](#)[FOR AUTHORS ▾](#)[CME >](#)

### ORIGINAL ARTICLE

## Randomized Trial of Peanut Consumption in Infants at Risk for Peanut Allergy

George Du Toit, M.B., B.Ch., Graham Roberts, D.M., Peter H. Sayre, M.D., Ph.D., Henry T. Bahnson, M.P.H., Suzana Radulovic, M.D., Alexandra F. Santos, M.D., Helen A. Brough, M.B., B.S., Deborah Phippard, Ph.D., Monica Basting, M.A., Mary Feeney, M.Sc., R.D., Victor Turcanu, M.D., Ph.D., Michelle L. Sever, M.S.P.H., Ph.D., Margarita Gomez Lorenzo, M.D., Marshall Plaut, M.D., and Gideon Lack, M.B., B.Ch. for the LEAP Study Team

N Engl J Med 2015; 372:803-813 | [February 26, 2015](#) | DOI: 10.1056/NEJMoa1414850

# Study features

**Design:** Longitudinal study  $\Rightarrow$  multiple measurements on each child over time.

**Outcome:** Presence of peanut allergy (skin prick + immunological markers) at 60 months of age

**Predictor of Interest:** Treatment consisting of feeding young children a peanut snack (or avoiding peanuts)

**Source(s) of correlation:** Multiple measurements on each child over time; not all individual characteristics (genetics, environment, etc.) can be measured



# Example 3 of study with correlated data

## THE LANCET

[Online First](#) [Current Issue](#) [All Issues](#) [Special Issues](#) [Multimedia](#) [Information for Authors](#)

All Content



Search

[Advanced Search](#)

[< Previous Article](#)

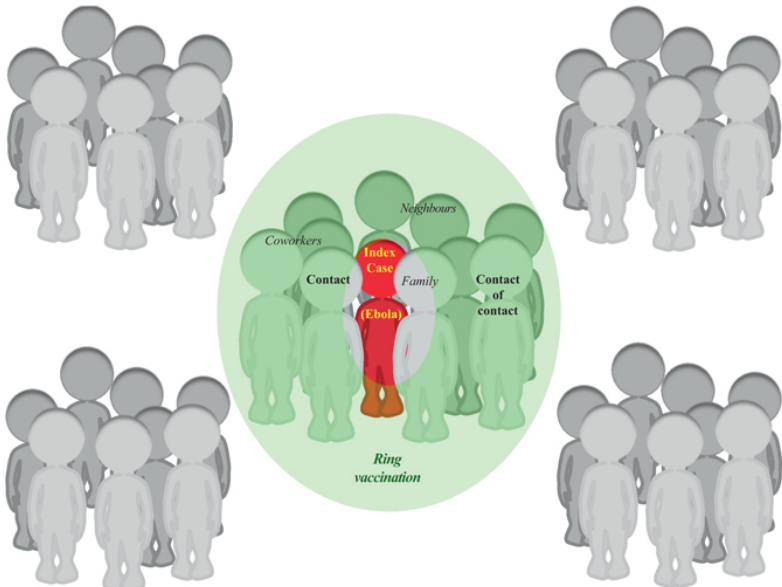
Volume 386, No. 9996, p857–866, 29 August 2015

[Next Article >](#)

Articles

### Efficacy and effectiveness of an rVSV-vectored vaccine expressing Ebola surface glycoprotein: interim results from the Guinea ring vaccination cluster-randomised trial

Ana Maria Henao-Restrepo, MD, Prof Ira M Longini, PhD, Prof Matthias Egger, MD, Natalie E Dean, PhD, Prof W John Edmunds, PhD, Anton Camacho, PhD, Miles W Carroll, PhD, Moussa Doumbia, MD, Bertrand Draguez, MD, Sophie Duraffour, PhD, Godwin Enwere, FWACP, Rebecca Grais, PhD, Stephan Gunther, MD, Stefanie Hossmann, MSc, Prof Mandy Kader Kondé, PhD, Souleymane Kone, MSc, Eeva Kuisma, PhD, Prof Myron M Levine, MD, Sema Mandal, MD, Gunnstein Norheim, PhD, Ximena Riveros, BSc, Aboubacar Soumah, MD, Sven Trelle, MD, Andrea S Vicari, PhD, Conall H Watson, MFPH, Sakoba Kéïta, MD, Dr Marie Paule Kieny, PhD<sup>✉</sup>, Prof John-Arne Røttingen, MD<sup>†</sup>



# Study features

**Design:** Cluster-randomized trial

**Outcome:** Infection with Ebola  $>$  10 days after vaccination

**Predictor of Interest:** Immediate or delayed vaccination

**Source(s) of correlation:** Individuals within a vaccination “ring” share a common environment and exposures.

# Example 4 of study with correlated data





## Spatial Statistics

Volume 9, August 2014, Pages 166–179

Revealing Intricacies in Spatial and Spatio-Temporal Data: Papers from the Spatial Statistics 2013 Conference



Exploration of the use of Bayesian modeling of gradients for censored spatiotemporal data from the *Deepwater Horizon* oil spill

Harrison Quick<sup>a</sup>, , Caroline Groth<sup>b</sup>, Sudipto Banerjee<sup>b</sup>,  , Bradley P. Carlin<sup>b</sup>, Mark R. Stenzel<sup>c</sup>, Patricia A. Stewart<sup>d</sup>, Dale P. Sandler<sup>e</sup>, Lawrence S. Engel<sup>f</sup>, Richard K. Kwok<sup>e</sup>

# Study features

**Design:** Observational study

**Outcome:** Pollution levels related to Deepwater Horizon oil spill

**Predictor of Interest:** Proximity to spill in space and time

**Source(s) of correlation:** Observations “nearby” in space/time are more similar than those far away.

# Course overview: The Four Skills

This course is about **how to handle correlated data**. To do this appropriately, you will need to be adept at four things:

- ➊ **Recognition:** Identify situations where data may be correlated.
- ➋ **Description:** Visualize and summarize correlated data in an informative way.
- ➌ **Modeling/estimation:** Choose and fit statistical models which account for correlation and respond to the scientific question.
- ➍ **Inference/interpretation:** Correctly interpret the results of analyses; understand their assumptions and the potential consequences if they are violated.

# Skill 1: Recognition

We have seen how different kinds of sampling/design can lead to correlated observations:

- Multiple (within-subject) measurements over time
- Measurements clustered in space/location
- Measurements on units which mutually influence each other (eg. social network)

Focus of this course will be primarily on this first study design, i.e. analysis of data from **longitudinal studies**, but the skills you learn are applicable to other kinds of clustered data.

# Why longitudinal studies?

Longitudinal studies are commonly employed in the sciences for several reasons:

- ❶ **Convenience:** Taking multiple measurements on each individual easier (= cheaper!) than recruiting additional subjects.
- ❷ **Addressing scientific questions:** Effects over time or within subjects.
- ❸ **Statistical efficiency:** Each subject acts as their own “control”, may be able to estimate effects more precisely.



## Skill 2: Description

- Working with correlated data: long format vs. wide format, a few helpful software tricks
- Summary statistics: outcome “flavors”, summarizing longitudinal trajectories
- Plotting correlated data: guiding principles, scatterplot smoothers
- Investigating correlation structures: scatterplot and empirical correlation matrices

## Skill 3: Modeling and estimation

We will learn about two main types of **regression models for correlated data**:

Generalized Estimating Equations (GEE)  
and Generalized Linear Mixed Models (GLMM)

- Model choice
- Assumptions
- Relevant software commands in R and SAS.

## Skill 4: Inference and interpretation

- Interpreting software output for GLMMs and GEE
- Assumptions: What might happen if you're wrong?
- Diagnostics: How do you check if assumptions are violated?
- Sample size: What  $N$  is “large enough” to trust your results?

# Next time...

- A slightly more formal introduction to correlation, and...
- Assignment 0