UNIVERSITY OF KENTUCKY
COLLEGE OF PUBLIC HEALTH

Course Syllabus
CPH 636-001 Data Mining in Public Health
Spring 2021

This is a fully online course which will have asynchronous and synchronous components. Lectures will be delivered asynchronously. You will be given access to lecture notes, plus accompanying video or audio recordings, on a regular basis. You should read the lecture notes and listen to the recordings within one week of their availability. You will be required to have at least seven synchronous meetings with me during the semester, and you may choose to have more. These synchronous meetings can be at mutually agreeable times, but I ask that you not commit to other recurring activities on Tuesdays and Thursdays from 3:30 to 4:45 p.m. That way, these times can be used as needed for synchronous meetings with me and/or your classmates.

Contact information

Instructor: Dr. Richard Charnigo

Telephone: Because of COVID-19, I am not in the office this semester. E-mail is the best way to reach me.

E-mail: richard.charnigo@uky.edu Please include “CPH 636” in the subject line. I try to respond to e-mail (when a response is called for) within one business day.

Office Hours: In addition to required synchronous meetings, I am available for extra Zoom sessions {uky.zoom.us} with individual students, or groups of students, upon reasonable request. Please e-mail me to schedule.

Course description

This course concerns statistical techniques for and practical issues associated with the exploration of large public health data sets, the development of models from such data sets, and the effective communication of one’s findings.

Course rationale

This three-credit-hour course provides methodological tools and conceptual insights which are not ordinarily part of core biostatistics courses. These tools and insights are relevant to discovering new relationships in substantive scientific contexts. Other content in the textbook, which the student may choose to self-study following completion of this course, has ties to contemporary research problems in statistics.

Course delivery

A broad overview of course delivery appears at the top of this page. Details regarding synchronous meetings appear below in the context of course requirements and learner evaluation.

Course prerequisites

BST 600 and CPH 535, or consent of instructor.

Course Objectives/Student Learning Outcomes and related UKCPH Competencies

The first column below shows program-level student learning outcomes and related UKCPH competencies for the Ph.D. program in Epidemiology and Biostatistics. Course objectives (i.e., course-specific student learning outcomes) are shown in the second column, along with parenthetical indications of competencies to which they may correspond. The course objectives in the second column also apply to students in other master's and doctoral programs; however, mappings to program-level student learning outcomes and related competencies (if any) would necessarily differ for students in other programs.

Program Level Outcomes	Course/Student Learning Outcomes
1. Demonstrate systems thinking using epidemiology theory and concepts and through data collection, analysis, interpretation, evidence-based reasoning. 1a. Evaluate the strengths and limitations of epidemiologic reports. 1b. Understand the principles of epidemiologic study design and be able to calculate the appropriate epidemiologic measures for most typical designs. 1c. Understand the principles of chronic and infectious disease epidemiology. 2. Analyze data and research methods using biostatistics theory and concepts. 2a. Draw appropriate inferences from data. 2b. Demonstrate an understanding of concepts of probability and statistical inference as they apply to problems in public health. 2c. Become proficient at and be able to evaluate the strengths and limitations of advanced designs including multivariate linear models, generalized linear models, longitudinal models, mixed effects models, and survival models both parametric and nonparametric. 3. Integrate biostatistics and epidemiological concepts in study design, implementation, analysis, and results interpretation from databases in the public health and medical research domains 3a. Understand the interface between biostatistics and epidemiology. 3b. Demonstrate advanced proficiency to apply concepts and methods from these disciplines jointly. 3c. Demonstrate proficiency in using computing tools commonly encountered in epidemiology and biostatistics. 3d. Demonstrate an understanding of research methods used in epidemiology and biostatistics. 4. Communicate inter-professionally regarding study management processes, problem conceptualization, ethics and core public health knowledge. 4a. Demonstrate the ability to review and critically evaluate the literature in a substantive area of research, be able to identify gaps in knowledge and be able to formulate original research hypotheses or statements. 4b. Communicate research results orally and in writing to lay and professional audiences. 4c. Demonstrate knowledge of the public health system in the commonwealth and the country.	I. Articulate the challenges associated with the acquisition and analysis of large public health data sets. (2b, 3d) II. Judiciously apply linear regression methodology to problems in public health. (2a, 2b, 3c, 3d) III. Judiciously apply linear classification methodology, including logistic regression, to problems in public health. (2a, 2b, 3c, 3d) IV. Employ classification and regression trees to analyze large public health data sets. (2a, 2b, 3c, 3d) V. Employ neural networks and nearest neighbor methods to analyze large public health data sets. (2a, 2b, 3c, 3d) VI. Articulate the strengths and weaknesses of various supervised learning techniques. (2b, 2c, 3d) VII. Apply unsupervised learning techniques to problems in public health. (2a, 2b, 3c, 3d) The team projects, along with the individual oral presentation and final written report, will fulfill competency 4b. Competencies 4a and 4c may also be fulfilled, to some degree, by virtue of completing the individual oral presentation and final written report.

Textbooks and Other Materials

1. The following textbook is required:

Hastie, Tibshirani, and Friedman (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd edition. Springer, New York.

From {http://statweb.stanford.edu/~tibs/ElemStatLearn}, you may freely and legally download an electronic copy. Please do not try to make a hard copy using a University printer.

2. You will need access to statistical software which can execute the various data mining methods we discuss. The free programming environment R {https://www.r-project.org} is one option, and the programming environment SAS – to which UKY has an institutional subscription – is another. These are two options with which I can assist based on my own knowledge and experience. You are welcome to use other programming environments (Java, Python, C++), but I may not be able to assist. You may use – possibly after modification – existing functions, procedures, macros, and/or syntax published by others. When you do so, you should so acknowledge.

3. Materials for this course will be posted on Canvas. Please check for updates frequently.

Course requirements and learner evaluation

Team Projects: There will be three team projects for you to complete. They will be due at 11:59 p.m. (Lexington KY local time) on the Tuesdays of 02 March, 23 March, and 13 April.

For each team project, I will propose teams, each of which will consist of three or four enrolled students. You are free to modify my proposed teams, as long as the resulting teams still consist of three or four enrolled students. You may also discuss the team projects with anyone you wish, including members of otherteams, but each team must prepare and is responsible for its own submitted work.

Each team member must contribute to the problem-solving process on each item within a project and be able to defend the response which is submitted for grading; a “divide and conquer” approach, whereby the different team members assume sole responsibility for various subsets of items, is not acceptable. (This is not to say that team members need to write redundant R or SAS code; however, please do not simply assign item 1 to person A, item 2 to person B, and so forth.) Please acknowledge any classmates whose insights informed your work, if those classmates are not members of your team.

For each of the three team projects, each team is required to have two Zoom meetings with me: once before the project is formally submitted for grading (to ask questions about items and/or request my comments on draft solutions) and once afterward (to receive my comments on submitted solutions and a grade). The first meeting should be within 14 days of the due date, and the second meeting should be within 7 days following the due date. Plan on 40-60 minutes for the first meeting and 20-30 minutes for the second meeting.

Please note, you are not limited to these Zoom meetings; I am available for other Zoom meetings upon reasonable request, to discuss team projects, answer questions about lecture notes or recordings, or attend to other course-related matters.

Team projects must be typed and submitted by e-mail as PDF files. Please include “CPH 636” in the subject line of the e-mail.

Individual Oral Presentation and Final Written Report: You will choose a data set that has not been considered in any of the lectures or team projects posted prior to 11:59 p.m. on Tuesday 02 March. The sample size must be at least 200, there must be at least 6 distinct variables, at least 3 of the variables must be treatable as continuous, and the data must be relatable to human health. You are welcome to use a data set that you are analyzing for your own research, if the data set meets the above requirements, if such use does not violate any conditions from a data use agreement or the Institutional Review Board, and if your research supervisor agrees. Otherwise, you may use a suitable data set from an Internet repository.

To ensure that you have ample time to prepare the oral presentation and final written report, please choose your data set and notify me (e-mail is fine) by 11:59 p.m. on Tuesday 09 March. No more than two students may work with the same data set, and data sets will be “reserved” on a first- come first-served basis. You are also required to schedule at least one Zoom meeting with me to discuss your progress with your data set, sometime between Thursday 18 March and Thursday 08 April. Plan on 20-30 minutes for this Zoom meeting.

Later I will furnish more detailed guidelines on what is expected for the oral presentation and final written report, but I will mention now that the oral presentation should be 20-30 minutes. The oral presentation and final written report are individual activities, but you are permitted to seek limited input or advice from classmates in much the same way that a faculty member might approach colleagues on a research problem. Roughly speaking, the level of classmate input or advice deemed permissible is such that, if you were to honestly assess the value added by your classmates, that value would merit a polite acknowledgment at the end of the presentation or report but would not warrant co-authorship. If you need more specific guidance, please ask.

You will record your oral presentation using Zoom and provide me with the hyperlink by 11:59 p.m. on Tuesday 27 April. I will view your oral presentation, and the hyperlink will be shared with other members of the class on Canvas. This means that other students may view your oral presentation, and you may view theirs. Any use of another student’s oral presentation besides your own viewing requires that student’s prior permission.

Because of the number of students in the class, the altered modality due to COVID-19, and University regulations on dead week, this semester I will not impose any formal requirement for you to listen to or comment on other students’ oral presentations. However, you are encouraged to view at least a few of them.

Final written reports will be due on Tuesday 11 May at 11:59 p.m., unless the registrar’s final exam schedule (not yet available as I prepare this document) requires a later deadline. Final written reports must be typed and submitted by e-mail as PDF files. Please include “CPH 636” in the subject line of the e-mail.

Grading Components: Each of the three team projects will count for 75 points (65 points for content and 10 points for the two required synchronous meetings), the oral presentation will be worth 75 points (65 points for content and 10 points for the required synchronous meeting), and the final written report will be valued at 100 points. Some extra credit may be available. The thresholds will be 90%, 75%, and 60% for “A”, “B”, and “C” letter grades respectively, although an outstanding final written report may result in the higher letter grade if you just barely miss the threshold. There will not be a “D” letter grade because this is a graduate course. Actually, no one should earn less than a “B”. (I do not promise that outcome, but I expect it.) This is an elective course and is meant to strengthen your data analysis and communication skills, not stress you out.

Instructor expectations

1. I expect you to listen to all video/audio commentaries (and review the corresponding lecture notes) within one week of their being posted on Canvas.

2. Please check the e-mail address under which you registered for the course regularly. As a courtesy, I will add alternate e-mail addresses to my mailing list upon request. You are responsible for all material and announcements conveyed by e-mail; a full mailbox or bouncing of messages by your e-mail provider does not remove this responsibility.

3. You are encouraged to ask questions by e-mail. Besides the required Zoom meetings, you may request appointments with me on Zoom. Prior permission from me (and from any other attendees, if applicable) is required for a student to initiate recording of a Zoom meeting.

4. Grading of written work will be based primarily on appropriateness of concept or methodology, technical accuracy or logic, and soundness of conclusions. I may also consider clarity, succinctness, and adherence to appropriate conventions of English language.

5. If you wish to appeal my grading, you may present an appeal in writing (by e-mail). However, this must be done within one week of the time my grading is conveyed to you.

6. The textbook is at a high mathematical level, and you are not expected to understand all of the details. You are also not expected to read, word-for-word, every section mentioned in the video/audio commentaries. Read what you need to read for better understanding. If you are not sure what to read, then ask me for advice.

Academic Policies

It is the student's responsibility to be informed concerning all regulations and procedures required by the program of study, College or the University. Students should become familiar with the Undergraduate Bulletin or the Graduate School Bulletin as appropriate. Academic disputes will be evaluated against these policies. This serves as formal notification of academic policies.

Students and faculty can locate the current University policies in the Syllabus section of their Canvas course. Please visit https://www.uky.edu/canvas/index.html and login with your linkblue ID.

Policies that are available include:

- Academic Integrity
- Accommodations Due to Disability
- Religious Observances
- Inclement Weather
- Excused Absences Policy
- Verification of Absences
- Student Resources

A hard copy of the policies will be provided by the Office of Academic Affairs upon request by the student.

Classroom Recording and Copyright Statement

The University of Kentucky Code of Student Conduct defines Invasion of Privacy as using electronic or other devices to make a photographic, audio, or video record of any person without their prior knowledge or consent when such a recording is likely to cause injury or distress.

Prior permission from me (and from any other attendees, if applicable) is required for a student to initiate recording of a Zoom meeting. I will not ordinarily initiate recording of Zoom meetings myself.

All video and audio recordings provided or initiated by me are not to be shared with those not enrolled in the class nor uploaded to other online environments. However, you may download them to a location accessible only to you.

Students with specific recording accommodations approved by the Disability Resource Center should present their official documentation to me.

Course Copyright

All original instructor-provided content for this course, which may include handouts, assignments, and lectures, is the intellectual property of the instructor. Students enrolled in the course this academic term may use the original instructor-provided content for their learning and completion of course requirements this term, but such content must not be reproduced nor sold. Students enrolled in the course this academic term are hereby granted permission to use original instructor-provided content for reasonable educational and professional purposes extending beyond this course and term, such as studying for a comprehensive or qualifying examination in a degree program, preparing for a professional or certification examination, or to assist in fulfilling responsibilities at a job or internship; other uses of original instructor-provided content require written permission from the instructor in advance.

As noted above, any use of another student’s recorded oral presentation besides your own viewing requires that student’s prior permission.

Attendance Policy

Attendance at required synchronous meetings is incorporated into the grade as described above.

Late work policy

Cases involving the following will be handled individually: excused absences (including religious observances), University-prescribed academic accommodations, and recommendations for special consideration from the office of an appropriate Dean or the Ombud.

Otherwise, the team projects, (recorded) oral presentation, and final written report will be accepted up to 24 hours past their respective deadlines without question or penalty. Submissions after 24 hours but within 48 hours will be accepted subject to a 25% penalty. Submissions after 48 hours will not be accepted.

Note: I understand that life is different with COVID-19. If you have difficulty keeping up with course requirements for a reason related to COVID-19 which does not fall under the University policy on excused absences, please let me know. I do not promise to agree to every request (for extension, accommodation, etc.), but I will try to be reasonable.

Course schedule and topics

This schedule is tentative and subject to change.

Week of 25 January: Chapter 1, Chapter 2 (Sections 1 – 3)
    Introduction, Terminology, Parametric versus nonparametric methods
Week of 01 February: Chapter 2 (Sections 4 – 6)
    Supervised learning, Curse of dimensionality
Week of 08 February: Chapter 2 (Sections 7 – 9)
    Regularization, Bias/variance tradeoff
Week of 15 February: Chapter 3 (Sections 1 – 3)
    Linear regression, Subset selection
Week of 22 February: Chapter 3 (Sections 4 – 6)
    Shrinkage, Derived directions
Week of 01 March: Chapter 4 (Sections 1 – 3)
    Linear and quadratic discriminant analysis
Week of 08 March: Chapter 4 (Section 4)
    Logistic regression
Week of 15 March: Chapter 7 (Sections 1 – 4)
    Bias/variance decomposition, Training/validation/testing data, Optimism
Week of 22 March: Chapter 7 (Sections 5 – 7, 10)
    Model complexity, Information criteria, Cross validation
Week of 29 March: Chapter 9 (Sections 1 – 2, 6)
    Regression/classification trees, Missing data
Week of 05 April: Chapter 11 (Sections 1 – 4)
    Neural networks
Week of 12 April: Chapter 13 (Sections 1 – 3)
    Nearest neighbors
Week of 19 April: Chapter 14 (Sections 1 – 3)
    Unsupervised learning, Cluster analysis
Week of 26 April: Chapter 10 (Sections 1, 7, 9)
    Boosting
Week of 03 May (Dead Days and Reading Days): Chapter 8 (Section 10)
    Stacking
Week of 10 May (Finals Week): No new topics
    Submit final written report

If our pace is faster than anticipated, then additional material may be drawn from Chapter 5 (Sections 1 – 2, 4 – 5) and/or Chapter 6 (Sections 1 – 3). If our pace is slower than anticipated, then material from Chapter 8 (Section 10) and/or Chapter 10 (Sections 1, 7, 9) may be omitted.

Other course-related information

If an unforeseen contingency arises that requires a new course policy, or if some clarification is warranted, then I will make an appropriate announcement.