QUEEN'S UNIVERSITY

Department of Public Health Sciences

Applied Statistical Learning for Health Data

Fall 2022

Course Instructors:

- Dr. Zihang Lu (course coordinator)
  Department of Public Health Sciences
  Email: zihang.lu@queensu.ca

- Dr. Wei Tu
  Department of Public Health Sciences & Canadian Cancer Trials Group
  Email: wei.tu@queensu.ca

Lecture Time and Location: 1:00pm – 4:00pm, Carruthers Room 311

Course Format: each week will include a 2-hour lecture and a 1-hour R lab session. The R lab session will cover exercises and problems using methods covered in the lecture.

Prerequisites: EPID 821 and EPID 822, or equivalent under the permission of the instructors.

Course Web Sites: A course web page at OnQ (https://onq.queensu.ca/) is available and used throughout the teaching. All course materials and announcements for this course will be posted there.

Course Description: Statistical learning refers to a set of methods that lie in the intersection of statistics and computer science for understanding the data. These methods can be roughly classified as supervised or unsupervised learning methods. The purpose of this course is to expose students to a variety of modern topics of statistical learning for health data. In this course, we will discuss several motivating problems in the healthcare domain, through which we will introduce fundamental statistical learning methods and their application to analyzing health data. The focus of this course will be the application and implementation of these statistical learning methods using large and real-world health data. Multiple case studies using health care data will be presented and discussed during the course. Previous programming experience for data analysis using R, SAS or other similar software is preferred but not necessary. Lecture examples will be given using R language.

Learning Objectives: This course is for students who are interested in learning about applied statistical learning methods and obtaining practical experience in the application of these methods to real-world data. This course does not intend to discuss in-depth theoretical and mathematical materials. Learning objectives include:

- Gaining an understanding of different types of statistical learning methods.
- Being able to implement these methods to health data using statistical software.
- Being able to choose appropriate methods for complex data problems and interpret the results.
- Being able to critically evaluate the use of basic statistical learning methods in public health literature.

Textbook: A majority of the courses materials will be from

- James, G.,Witten, D., Hastie, T. and Tibshirani, R., 2013. An introduction to statistical learning (Vol. 112, p. 18). New York: springer.

Reference Books: Following books will serve as reference materials.

- Hoyt, R. and Muenchen, R., 2019. Introduction to biomedical data science. Lulu. com.
- Malley, J.D., Malley, K.G. and Pajevic, S., 2011. Statistical learning for biomedical data. Cambridge University Press.
- Friedman, J., Hastie, T. and Tibshirani, R., 2001. The elements of statistical learning (Vol. 1, No. 10). New York: Springer series in statistics.
- Bishop, C.M., 2006. Pattern recognition and machine learning. springer.

Other Reading Materials:

- Gelman, A. and Vehtari, A., 2021. What are the most important statistical ideas of the past 50 years?. *Journal of the American Statistical Association*, *116*(536), pp.2087-2097.
- Wing, J. M., 2020. Ten Research Challenge Areas in Data Science. Harvard Data Science Review.
- Wiemken TL, Kelley RR. Machine Learning in Epidemiology and Health Outcomes Research. Annual review of public health. 2019 Oct 2;41:21-36.
- Donoho, D., 2017. 50 years of data science. Journal of Computational and Graphical Statistics, 26(4), pp.745-766.
- Jordan, M.I. and Mitchell, T.M., 2015. Machine learning: Trends, perspectives, and prospects. Science, 349(6245), pp.255-260.
- Breiman, L., 2001. Statistical modeling: The two cultures (with comments and a rejoinder by the author). Statistical science, 16(3), pp.199-231.

Evaluation: The final grade for this course will be based on the following evaluations: the grading will be based on three take-home problem sets (each accounting for 20%) plus a final take-home project (40%).

Please note:

- The minimum passing grade in Graduate School is 70% for this course.
- Late assignments without valid reasons will only receive 75% of marks if handed in before solutions are posted.
- There will be no make-up quizzes and homework assignments. Students who miss any of these evaluations for a VALID REASON (proof required) will have the percentages of the missed evaluations allocated to the remaining evaluations.
- Students who request accommodation should go through Queen's accessibility services and their Program Director (rather than their course instructors):

https://www.queensu.ca/studentwellness/accessibility-services

Tentative Weekly Schedule for Lectures:

Week 1: Overview of the Course (Lu)

This module provides an overview of the courses and introduces different types of health data and problems.

- Overview of the course structure
- Overview of common topics in health data research
- Overview of statistical learning methods
- Types and examples of health data

Week 2: Exploratory Data Analysis (Part 1) (Lu)

This module introduces some common methods to process the data prior to statistical analysis.

- Understand different types of data
- Data preparation and data cleaning
- Data visualization

Week 3: Exploratory Data Analysis (Part 2) (Lu)

This module introduces some common methods to perform exploratory data analysis.

- Dealing with missing data
- Dealing with outliers
- Reporting the summary statistics

Week 4: Linear Regression (Lu)

This module will introduce linear regression. We will discuss how to build, estimate and interpret simple and multiple linear regressions.

- Simple linear regression
- Multiple linear regression
- Estimating and interpreting the regression coefficients

Week 5: Linear Model Selection and Regularization (Lu)

This module will introduce linear regression for prediction problems. We will discuss how to build, estimate and interpret simple and multiple linear regressions.

- Subset selection and stepwise regression
- Shrinkage methods using ridge and lasso regression
- Consideration in high dimensional regression

Week 6: Non-linear Regression (Lu)

This module will introduce common techniques for regression analysis when the linearity assumption is not met.

- Polynomial regression
- Regression splines
- Smoothing splines
- Local regression
- Generalized additive models

Week 7: Disease Modelling via Classification (Part 1) (Tu)

This module will introduce model assessment approach and the classification (supervised learning) problem. Common approaches to perform classification will be introduced and discussed.

- Cross-validation
- Linear discriminant analysis
- Linear classification
- Logistic regression
- Naive Bayes

Week 8: Disease Modelling via Classification (Part 2) (Tu)

This module will continue to introduce more approaches to perform classification, focusing on tree-based methods, such as decision trees, random forest and support vector machines.

- Decision trees
- Random forests
- Support vector machines

Week 9: Methods for Dealing with High Dimensional Data (Part 1) (Tu)

This module will introduce common techniques used in high dimensional settings, focusing on variable selection (e.g. best subset selection), model selection and shrinkage methods (e.g. lasso). These methods are widely used to reduce the model complexity in high-dimensional settings.

- Variable selection problem
- Ridge and lasso
- Considerations in high dimensions

Week 10: Methods for Dealing with High Dimensional Data (Part 2) (Tu)

This module will continue to introduce common techniques used in high dimensional settings, focusing on dimension reduction techniques that are wildly used in high dimensional settings.

- Principal component analysis
- Correspondence analysis
- T-distributed Stochastic Neighbor Embedding

Week 11: Pattern discovery in Health Data (Tu)

This module will introduce several analytical tools for clustering (unsupervised learning) problems. Common approaches to performing cluster analysis will be introduced and discussed via a case study.

- Partitioning methods
- Hierarchical clustering
- Density-based clustering
- Model-based clustering

Week 12: Additional Topics and Review (Tu)

This module will briefly introduce some additional topics in statistical learning, and will also provide a review of topics discussed previously in this course.

Statement on Academic Integrity

The following statement on academic integrity builds on a definition approved by Senate and is designed to make students aware of the importance of the concept and the potential consequences of departing from the core values of academic integrity. **It is required that this statement be included on all course syllabi.** Instructors may also consider including this statement with each assignment.

Queen's students, faculty, administrators and staff all have responsibilities for upholding the fundamental values of academic integrity; honesty, trust, fairness, respect, responsibility and courage (see www.academicintegrity.org). These values are central to the building, nurturing and sustaining of an academic community in which all members of the community will thrive. Adherence to the values expressed through academic integrity forms a foundation for the "freedom of inquiry and exchange of ideas" essential to the intellectual life of the University (see the Senate Report on Principles and Priorities http://www.queensu.ca/secretariat/policies/senate/report-principles-and-priorities).

Students are responsible for familiarizing themselves with the regulations concerning academic integrity and for ensuring that their assignments and their behaviour conform to the principles of academic integrity. Information on academic integrity is available in the SGS Calendar (https://www.queensu.ca/sgs/graduate-calendar/academic-integrity-policy) and from the instructor of this course. Departures from academic integrity include plagiarism, use of unauthorized materials, facilitation, forgery, falsification and unauthorized use of intellectual property, and are antithetical to the development of an academic community at Queen's. Given the seriousness of these matters, actions which contravene the regulation on academic integrity carry sanctions that can range from a warning or the loss of grades on an assignment to the failure of a course to a requirement to withdraw from the university.

It is recommended that instructors add a paragraph here to explain issues of academic integrity that are particularly relevant to the course. For example:

- Plagiarism – including guides on how to use sources correctly. Possible example:
    - *Please note that we have had issues in the past with unintended plagiarism in this course. Regardless of how and where you retrieve information, the principles of academic integrity apply. Please visit these helpful websites to help you make sure that you are able to write things in your own words:*
- *https://www.queensu.ca/academicintegrity/students/avoiding-plagiarismcheating*
- *https://integrity.mit.edu/handbook/academic-writing/avoiding-plagiarism-paraphrasing*
- *http://writing.wisc.edu/Handbook/QPA_paraphrase.html*

- Groupwork – what level of collaboration is acceptable? Clearly state if there are things students must do alone. Possible example:
    - *You are permitted to work with a partner or in groups of 3 to encourage collaboration, cooperation, and collective learning on lab assignments. You are not permitted to share answers among large groups or as a tutorial group. You must work independently on quizzes and "pop questions".*

Statement on Copyright of Course Materials

The following statement is pre-loaded into courses created in onQ:

Course materials created by the course instructor, including all slides, presentations, handouts, tests, exams, and other similar course materials, are the intellectual property of the instructor. It is a departure from academic integrity to distribute, publicly post, sell or otherwise disseminate an instructor's course materials or to provide an instructor's course materials to anyone else for distribution, posting, sale or other means of dissemination, without the instructor's express consent. A student who engages in such conduct may be subject to penalty for a departure from academic integrity and may also face adverse legal consequences for infringement of intellectual property rights.

Privacy Statement for Instructors Who Use External Software in Their Course

This course makes use of [name of software or company] for xxxxxxxx. Be aware that by logging into the site, you will be leaving onQ, and accessing [the name of company's] website and [name of software application]. Your independent use of that site, beyond what is required for the course (for example, purchasing the company's products), is subject to [name of company's] terms of use and privacy policy. You are encouraged to review these documents, using the link(s) below, before using the site.

Links to the most common websites used by instructors are listed below:

- Crowdmark - https://crowdmark.com/privacy/queens/
- Pearson & Peer Scholar- http://www.pearsoncanada.ca/pearson-canada-at-a-glance/legal/privacy-statement
- Wiley - https://www.wiley.com/en-ca/privacy
- McGraw Hill - https://www.mheducation.ca/privacy/
- Turnitin - http://turnitin.com/en_us/about-us/privacy
- Rosetta Stone (formerly Tell Me More) - http://resources.rosettastone.com/CDN/us/agreements/US_Privacy_Policy-102513.pdf
- Coglab - https://coglab.cengage.com/info/privacy.shtml

Statement on Academic Accommodations for Students with Disabilities

Queen's University is committed to achieving full accessibility for people with disabilities. Part of this commitment includes arranging academic accommodations for students with disabilities to ensure they have an equitable opportunity to participate in all of their academic activities. The Senate Policy for Accommodations for Students with Disabilities was approved at Senate in November 2016 (see https://www.queensu.ca/secretariat/sites/webpublish.queensu.ca.uslcwww/files/files/policies/senateandtrustees/ACADACCOMMPOLICY2016.pdf). If you are a student with a disability and think you may need academic accommodations, you are strongly encouraged to contact the ***Queen's Student Accessibility Services (QSAS)*** *and register as early as possible. For more information, including important deadlines, please visit the QSAS website at: http://www.queensu.ca/studentwellness/accessibility-services/*

Statement on Academic Consideration for Students with Extenuating Circumstances

Queen's University is committed to providing academic consideration to students experiencing extenuating circumstances that are beyond their control and are interfering with their ability to complete academic requirements related to a course for a short period of time. The Senate Policy on Academic Consideration for Students in Extenuating Circumstances is available at http://www.queensu.ca/secretariat/sites/webpublish.queensu.ca.uslcwww/files/files/policies/senateandtrustees/Academic%20Considerations%20for%20Extenuating%20Circumstances%20Policy%20Final.pdf

Each Faculty has developed a protocol to provide a consistent and equitable approach in dealing with requests for academic consideration for students facing extenuating circumstances. SGS students can find the Academic consideration information at: https://www.queensu.ca/sgs/accommodation-and-academic-consideration.

If you need to request academic consideration for this course, you will be required to provide the name and email address of the instructor/coordinator. Please use the following:

Instructor/Coordinator		*Name:*
Instructor/Coordinator	*email*	*address:*

Statement on Use and Retention of Video Recording (if using)

Include the following statement in the course syllabus if you plan to record your synchronous (live) classes or meetings and make the recordings available to students in your class afterwards:

Synchronous (live) classes will be delivered in this course through a video conferencing platform supported by the University [MS Teams, Zoom]. Steps have been taken by the University to configure these platforms in a secure manner. Classes will be recorded with video and audio (and in some cases transcription) and will be made available to students in the course for the duration of the term. The recordings may capture your name, image or voice through the video and audio recordings. By attending these live classes, you are consenting to the collection of this information for the purposes of administering the class and associated coursework. If you are concerned about the collection of your name and other personal information in the class, please contact the course instructor to identify possible alternatives.

To learn more about how your personal information is collected, used and disclosed by Queen's University, please see the general Notice of Collection, Use and Disclosure of Personal Information.

Statements on Remote Proctoring (if using)

a) All students must be informed at the start of the course that the instructor will be using a remote proctoring tool for tests/exams. In addition, it is recommended that instructors also address the use of the chosen remote proctoring tool at the outset of the course, whether the proctoring will be live or recorded, and the importance of academic integrity with students. The course syllabus must therefore contain the following statements for students:

The final exam and some tests/quizzes in this course will use remote proctoring provided by a third-party, cloud-based service that enables the completion of a proctored exam or test from an off-campus location, through onQ or Elentra. This online proctoring solution was chosen as part of the approach to maintaining academic integrity in online assessment. Precise details about how remote proctoring will be used in this course can be found in the "Getting Started with Remote Proctoring" content module in onQ or will be provided by the instructor.

When writing tests/exams using remote proctoring, you are connecting to the third-party service. Queen's has conducted a privacy and security review of the services in accordance with Ontario's privacy legislation, and has entered into binding agreements with Examity/Proctortrack.

You should also take measures yourself to protect your information by keeping your NetID password and challenge questions private, closing all applications prior to starting an exam/test, and ensuring your device is updated and safeguarded against malware.

For more information about remote proctoring, see the Student FAQs on the OUR Exams resource page for remote proctoring:

http://www.queensu.ca/registrar/students/examinations/exams-office-services/remote-proctoring

b) The course syllabus must contain the following statement for students who require academic accommodations in their exams as authorized by QSAS:

To have your accommodations applied to a remote-proctored exam please follow the instructions for the course, as outlined on the QSAS website. Your exam accommodations, as authorized by your Letter of Accommodation, will be incorporated into your Examity/Proctortrack exam session. Please note that exam accommodations that are uploaded for a specific exam are only visible to students once they begin their exam in the Exam Portal.

c) Additional information related to academic integrity in the context of remotely-proctored exams should be included for courses where exams will be monitored by remote proctoring.

Departures from academic integrity include plagiarism, use of unauthorized materials or services, facilitation, forgery, falsification, unauthorized use of intellectual property, and collaboration, and are antithetical to the development of an academic community at Queen's. Given the seriousness of these matters, actions which contravene the regulation on academic integrity carry sanctions that can range from a warning or the loss of grades on an assignment to the failure of a course to a requirement to withdraw from the University. In the case of online exams, impersonating another student, copying from another student, making information available to another student about the exam questions or possible answers, communicating with another person during an exam or about an exam during the exam window, or accessing unauthorized materials, including smart devices, are actions in contravention of academic integrity.