COURSE COLOUR CODE: MOCHA


## MASTER OF BIOTECHNOLOGY PROGRAM

*Compulsory Course Component*

## BTC1877H

# DATA SCIENCE IN HEALTH II

Nicholas Mitsakakis

Fall Term, 2022

# MASTER OF BIOTECHNOLOGY

## UNIVERSITY OF TORONTO MISSISSAUGA

## BTC1877H – Data Science in Health II

## Course Outline (Fall, 2022)

| | |
|---|---|
| Class Location: | Kaneff Centre, Room 112 (KN-112) |
| Class Times: | Wednesdays, 14-Sep to 14-Dec, 6:00-9:00PM |
| Instructor: | **Prof. Nicholas Mitsakakis** |
| Office Location: | N/A |
| Office Hours: | TBC |
| Contact: | n.mitsakakis@theta.utoronto.ca |

### Course Description

This course will build upon **Data Science in Health I** (BTC1859H). Themes continued from that course include the application of statistical analysis to problems rooted in biology and healthcare related applications. There will be more tutorials centered around "crap detection" and deriving insight in short scenarios in a biological or clinical context. To think creatively about new approaches to analysing data, you must first be able to think critically about such analysis. The major project will be a complex problem set in clinical or biological application.

The course will also deepen your analytical repertoire. You will learn more advanced tools in the context of machine learning. Some of these will be relevant to diagnostic applications in start-ups and research while others may be relevant in predictive activities in a health context. Students are expected to use appropriate tests in R and as appropriate, for larger models, obtain and validate R routines from public sources in order to build larger models.

You will also be expected to use your knowledge of biology to help guide your choices in how to tackle a data set. This means understanding both the biological system and the literature behind as it impacts how you analyse the problem.

In addition, you will also be expected to have the confidence and comfort to learn new analytical methods on your own, that you deem best suited for the analysis you wish to conduct. Independent learning is the core of graduate training and professional life.

### Course Objectives

At the conclusion of this course, students should be able to—

1) Understand and use methods for survival analysis;
2) Use advanced machine learning tools;
3) Train and apply neural networks for prediction;
4) Identify and apply the correct statistical tests for a given problem;
5) Implement the above in the programming language, R;

6) Understand the assumptions behind a statistical-tests for specific clinical and biological applications;
7) Critique the use of statistical tests for a given data set in a health context; and
8) Tackle novel data problems in data analysis where the appropriate statistical technique may not be known.

## Marking Scheme

The breakdown of the grade for the course will be as follows—

| | |
|---|---|
| Class Participation | 15% |
| Individual Homework Assignment 1 | 7½% |
| Individual Homework Assignment 2 | 7½% |
| Kobayashi Maru | 20% |
| Team Project | 25% |
| Final Exam | 25% |
| **TOTAL** | **100%** |

The following marking scheme is in effect for the **Individual Homework Assignments**—

| | |
|---|---|
| Analysis | 30% |
| Code | 20% |
| Interpretation | 30% |
| Writing and presentation | 20% |
| **TOTAL** | **100%** |

The following marking scheme applies for the **Team Project**—

| | |
|---|---|
| Presentation | 35% |
| Report | 65% |
| **TOTAL** | **100%** |

Marking scheme for the report is as follows—

| | |
|---|---|
| Analysis | 30% |
| Code | 20% |
| Interpretation | 25% |
| Writing and presentation | 25% |
| **TOTAL** | **100%** |

## Participation & Online Conduct of Classes

Participation mark is based on attendance and the quality of on-line participation. The quality of class participation includes appropriate preparation for the material and insightful questions or comments in class discussion. Students are required to prepare properly for each class by carefully studying the pre-class reading material that is provided in the course syllabus. Occasionally, quizzes may be handed in and students will have to take them. Marks of these quizzes will count as part of the participation

grade. Some of the tutorials will also include problems that are to be solved by the students during the tutorial time, and the solutions will be graded and count as part of the participation grade.

Students in an online course will login as requested by the instructor. Maintaining a professional appearance and attire throughout the duration of the online classes is required.

## Individual Homework Assignments

Two assignments will be given where students will need to apply some of the methods they will have learned. These assignments are assigned at the individual level and collaboration among the students as they work on these assignments is prohibited. Students will be given about 2 weeks to complete each assignment.

## Team Project

Students will need to analyse a data set (case study), to create a report and to present the analysis in the class. Presentations will take place on **7-Dec**. Each team is to present for 12 minutes. All team members must have equal presentation time. You should introduce the problem, your methods and analysis. The focus of your talk is on your analysis. If you cover material too quickly or you are cryptic, you will be penalised in your presentation score. Your team's ability to answer questions will be part of your presentation score.

## Further Recommendations for Team Project Presentations

To increase legibility font size should be no less than 18-point font (if using PowerPoint) and a limit of one figure per slide. To control presentation pace, at least one minute must be spent reviewing each slide, and it cannot be cluttered with content, which means the slide deck cannot exceed 12 slides (excluding the title slide). References should be at the bottom of each slide, not at the end of the slide deck.

## Kobayashi Maru

The **Kobayashi Maru** is an individual assessment of your ability to show insight into a problem in data analytics and your understanding of the foundations of the tests you apply. This is intended to be a very challenging situation.

Students will all receive the same problem-set based on real world data. Students will give a 5-minute slide presentation showing their most insightful finding(s) in the data set. There is a 5-slide limit on content—at least one minute must be spent on each slide. Each slide needs to contain, at most, one figure. The question period will be 15 minutes.

The question period will be an oral exam that will cover the student's presentation and any other questions regarding statistical analysis the judges deem fit to pose.

Students need to review statistical concepts covered in **Data Science in Health I** and be ready to answer any questions, which may go beyond the immediate analysis in their presentation. Examples of questions include: "What are the assumptions of your test? What are related tests? Why did you choose this method as opposed to others? What kind of confounding considerations do you need be vigilant of?"

Note that you can use any method you deem appropriate— including ones we have not learned in the courses—**but be prepared to answer questions on how the selected method works and why you chose it.**

Teamwork is **not** allowed, as this is an individual assignment. Your performance will be partly judged based on your insightfulness compared to the work done by your peers. Try to be sound, but also creative and unique.

Students will present one at a time. Only one student will be permitted in the "room" at a time and will only present to the judges. A schedule will be made available.

The Kobayashi Maru will be conducted over 2 class sessions. Students' performance will be compared to peer performance of the same day. Thus, students presenting on the 2nd day of this assessment will be compared to students who also presented on the 2nd day.

## Final Exam

The final exam will be up to 2 hours in duration and will involve short-answer questions. Material in the assigned readings, student presentations, guest lecturers, tutorials and all other material covered in class are all relevant material that could be tested for on the exam. The exam will be closed book. The exam will be conducted online, with more details to follow. **The date and time of the final exam will be Wednesday, 14-Dec, 6:00-8:00PM.**

## Course Materials

The materials in the slides and lecture notes will be the main resource for this course. The following textbooks will be used (not exclusively) during different sessions of the course:

1) Peter Dalgaard, *Introductory Statistics with R*, (2nd ed.), Springer (Lecture 1).
2) James, G., Witten, D., Hastie, T. & Tibshirani, R. (2013), *An introduction to statistical learning* (Vol. 112, p.18). New York: Springer (Lectures 4, 5, 9, 10).
3) van Buuren, S. (2018). Flexible Imputation of Missing Data (2nd ed.). Chapman & Hall/CRC. https://doi.org/10.1201/9780429492259 (Lecture 8).

Occasionally, other material may be provided for reading prior to the lectures.

## Software

The freely available statistical software R will be used for this course. Additionally the programming editor/environment RStudio will be used. Both R and RStudio can be downloaded freely from the Internet, from the following sites—

- R: https://www.r-project.org
- RStudio: http://www.rstudio.com

After installing R and RStudio, you should see icons in your **Applications** folder on OS X or **All Apps** in Windows 10. Starting RStudio (the editor) also starts an R session, so you do not need to run R directly.

**The students are required to download and install both R and RStudio prior to the first lecture/tutorial.**

# SCHEDULE OF ACTIVITIES

| Unit | Date | Topic | Assignment |
|---|---|---|---|
| 1 | 14-Sep | **Introduction to Survival Analysis:** Censoring, Kaplan-Meier curves, log-rank test, parametric models, Cox regression model, proportional hazard assumption | ○ **Tutorial:** plotting and analysis functions in R, survival R package<br>○ **Reading:** Textbook 1 (Ch 14)<br>○ **Assignment 1** is handed out (see Assignment description for due date) |
| 2 | 21-Sep | **Critical Thinking with Biostatistics Part I: Misleading Analysis (guest lecture by Dr. Jayson Parker)** Students will work on some in-class problem sets that will provide the opportunity to further practice critical thinking in conducting biostatistics. Expectations in working in industry will also be reviewed in the same context. Tackling problems outside one's comfort zone for analysis will also be discussed. | ○ **Tutorial:** Problems will be given and discussed<br>○ **Reading:** N/A |
| 3 | 28-Sep | **Critical Thinking with Biostatistics Part II: Small Data (guest lecture by Dr. Jayson Parker)** Similar format to Part I for this topic. While big data is the topic of much discussion, this class will introduce the problems presented by "small data". Small data is just as much part of professional life as large data sets, but requires a different approach. | ○ **Tutorial:** Problems will be given and discussed<br>○ **Reading:** N/A |
| 4 | 5-Oct | **Machine Learning, Part I:** Basic concepts, training vs. test error, overfitting, resampling methods, regularization in linear regression. | ○ **Tutorial:** ML functions in R, examples and exercises<br>○ **Reading:** Textbook 2 (p. 26-36, 175-190, 219-221) |
| | 12-Oct | **Reading week – *No classes*** | |
| 5 | 19-Oct | **Machine Learning, Part II:** Methods for classification, KNN, classification trees, random forests and other ensemble methods. | ○ **Tutorial:** ML functions in R, examples and exercises<br>○ **Reading:** Textbook 2 (p. 37-42, 303-316)<br>○ **Assignment 2** is handed out (see Assignment instructions for due date) |
| 6 | 26-Oct | **Kobayashi Maru, Part 1** | |
| 7 | 2-Nov | **Kobayashi Maru, Part 2** | |

| Unit | Date | Topic | Assignment |
|---|---|---|---|
| 8 | 9-Nov | **Methods of imputation of missing data:** Multiple imputation, chained equations | ○ **Tutorial:** examples of multiple imputation in R<br>○ **Reading:** Textbook 3 (pages TBA)<br>○ **Team Project:** data set and description is handed out |
| 9 | 16-Nov | **Unsupervised Learning:** K-means clustering, hierarchical clustering, methods of assessing clustering, principal component analysis (PCA) for dimension reduction | ○ **Tutorial:** Examples using R functions and biological data<br>○ **Reading:** Textbook 2 (p. 373-399) |
| 10 | 23-Nov | **Support Vector Machines:** linear SVM, soft margins, Support Vector Classifier, kernels | ○ **Tutorial:** examples of SVM in R<br>○ **Reading:** Textbook 2 (p. 337-355) |
| 11 | 30-Nov | **Introductions to Neural Networks:** Biological analogy, basic architectures, back propagation, gradient descent, introduction to Deep Learning | ○ **Tutorial:** examples of neural network training in R<br>○ **Reading:** TBD |
| 12 | 7-Dec | **Team Project Presentations:** Team presentations of 12 minutes each; Question period following each presentations | **Team Project Report is due** |

## Procedures & Rules

**MISSED TEST(S)/FINAL EXAM:** A student that misses a test due to illness must submit a completed University of Toronto Student Medical Certificate (available at: http://www.utm.utoronto.ca/registrar/sites/files/registrar/public/shared/pdfs/medcert_web.pdf) to the Instructor or Program Office (KN-209). Only the University of Toronto Student Medical Certificate will be accepted in support of petitions that cite illness as the reason for the request. Documentation concerning physician examinations must show that the physician was consulted on the day of the test date or immediately after, i.e. the next day. A statement from a physician that merely confirms a report of illness and/or disability made by the student is not acceptable. Documentation citing non-essential, preplanned medical procedures will not be acceptable. All documents must be originals and must be presented in person with a valid UofT student card within 72 hours of missing the test. Beyond 72 hours from the test date, further documentation of continued illness or disability will be required from a physician.

A student that misses a test due to domestic tragedy, at the discretion of the instructor, must provide acceptable documentation validating the explanation for absence.
If a test is missed and the student does not provide acceptable documentation validating the explanation for absence, a grade of “0” may be assigned at the instructor’s discretion.

If a test is missed and validating documentation is accepted the students are expected to write a make-up test. Students must contact the instructor immediately by phone or email to make arrangements.

**LATE ASSIGNMENTS:** Late assignments will be assigned a late penalty. We recognize that there may be valid reasons for late assignments. In order for these reasons to be accepted without penalty, supporting documentation will often be required. Extensions are given at the discretion of the instructor and require supporting documentation. In all cases, be sure to contact the Instructor before the assignment due date. The Instructor will ask you to report in writing your reasons for lateness and to state a revised due date.

Unless there is a previous communication with the instructor and specific permission has been obtained, a deduction of 10% marks per day will be applied for late work, with a maximum of 4 days. After that submissions will not be acceptable.

Instructors may not grant extensions beyond SGS course deadlines. Such considerations must be negotiated between students, instructors, the program director and SGS.

**ACADEMIC MISCONDUCT:** Students should note that copying, plagiarizing, or other forms of academic misconduct will not be tolerated. Any student caught engaging in such activities will be subject to academic discipline ranging from a mark of zero on the assignment, test or examination to dismissal from the university as outlined in the School of Graduate Studies academic handbook. Any student abetting or otherwise assisting in such misconduct will also be subject to academic penalties.

Students agree that by taking this course all required papers may be subject to submission for textual similarity review to Turnitin.com for the detection of plagiarism. All submitted papers will be included as source documents in the Turnitin.com reference database solely for the purpose of detecting plagiarism of such papers. The terms that apply to the University's use of the Turnitin.com service are described on the Turnitin.com web site.

## Communication

### LOGGING IN TO YOUR QUERCUS COURSE WEBSITE

Like many other courses, BTC1877H uses Quercus for its course website. To access the BTC1877H website, or any other Quercus-based course website, go to the UofT portal login page at: https://q.utoronto.ca and log in using your UTORid and password. Once you have logged in to the portal using your UTORid and password, look under the **Courses** menu item, where you'll find the link to the BTC1877H course website along with the link to all your other Quercus-based courses.

### E-MAIL COMMUNICATION WITH THE COURSE INSTRUCTOR

At times, the course instructor may decide to send out important course information by e-mail. To that end, all UofT students are required to have a valid UofT e-mail address. You are responsible for ensuring that your UofT e-mail address is set up AND properly entered in the ROSI system.

**Forwarding** your utoronto.ca e-mail to a Hotmail, Gmail, Yahoo or other type of e-mail account is not advisable. In some cases, messages from utoronto.ca addresses sent to Hotmail, Gmail or Yahoo accounts are filtered as junk mail, which means that e-mails from your course instructor may end up in your spam or junk mail folder.

You are responsible for:

1) Ensuring you have a valid UofT e-mail address, properly entered in the ROSI system
2) Checking your UofT e-mail account on a regular basis.