*Machine Learning for Precision Public Health Workshop #1 –*
*A conversation with thought leaders*

# Summary

5th of February, 2019, 1:15 – 2:45 PM
Latham Hall, University of British (UBC) Medical Student & Alumni Centre

## Overview of the workshop

This workshop was the first session of the special seminar series Machine Learning for Precision Public Health, organized by the BC Centre for Disease Control (BCCDC). Led by Dr. Jennifer Gardy, a BCCDC/UBC researcher who recently joined the Gates Foundation, and Tom Schenk, director of Smart Cities at KPMG and former chief data officer at the City of Chicago, this workshop aimed to:

- Provoke thoughts on potential opportunities to apply data science approaches to participant's own public health work;
- Facilitate discussions around the challenges of launching such work and possible solutions to address them; and
- Provide an opportunity for participants to learn from experts how to apply data science approaches in public health in the real-world setting

A total of 40 participants were registered for the workshop, including academic researchers, university students, clinical fellows, epidemiologists, statisticians, and public health practitioners from the BCCDC, UBC, Regional Health Authorities, BC Centre for Substance Use, Rick Hansen Institute, BC Cancer Research Centre, and many other organizations.

Dr. Gardy and Schenk opened the workshop with examples of health projects that could benefit from data science approaches based on their own experience, including:

- Incorporating tuberculosis (TB) genomics data to better trace and manage outbreaks
- Extracting lab reports using natural language automatic processing
- Applying machine learning algorithms to administrative data to predict risks for specific health outcomes

Workshop participants were then divided into two groups to brainstorm more project ideas, identify opportunities and challenges, and discuss possible actions to take. The workshop concluded with participants rejoining as a whole group, exchanging the key points of discussions and sharing their final thoughts and questions around data science and public health.

# Key points of discussions

## What is a data science project?

- It needs to have some automation in the data input and out processes
- It is usually prediction-focused and helps inform future actions
- It does not necessarily focus on causation
- It is usually problem-based; the approach will vary depending on the types of data available and ethical considerations on how the data is used.

## Ideas of a data science project in the health context

- To predict the impact of climate changes on population health
- To help detect cancer in an early stage through applying automatic classifications in histopathological diagnoses
- To predict clusters (hot spots) of cases of an infectious disease for timely management
- To create a scoring system to predict mortality risks for early interventions
- To identify populations to prioritize cancer screening
- To understand patterns of health care utilization (e.g., emergency room visits, hospital readmissions) for specific health outcomes

## What opportunities may data science bring to the health field?

- Building a community which brings together people with different expertise
- Helping identify what to prioritize and increase efficiencies in public health and clinical services
- Demonstrating the benefits and impacts of data on improving public health

## What are potential roadblocks to launching data science projects in public health?

- Data access which usually takes a long time to gain and sometimes involves a complicated bureaucratic process to navigate different data stewards
- Lack of resources, both human and financial
- Privacy, as regulated by the legislation and/or concerned by the public on collection and use of personal data
- Politics (e.g., lack of support due to the mismatched priorities of the government and funder; short project timelines given the nature of political terms, etc.)
- Funding timelines which influence the project scope and sometimes make it difficult to explore alternative methods
- No existing standards for best practices in applying machine learning in the public health context
- Limited knowledge of the types and amount of data needed to launch such projects
- Not enough connections and collaborations between subject matter experts and technical experts to develop appropriate prediction algorithms

- Confusion between correlation and causation in applying data science approaches to a health project
- Engagement of stakeholders who might not be comfortable with machine learning methods
- Difficulty in verifying the accuracy and quality of data sometimes, and managing missing data
- Lack of understanding and/or transparency in machine learning modelling which may seem like a black box to some, and hence lack of trust in the prediction outcomes
- Uncertainty around external validity when applying an existing machine learning algorithm to a different context

## What are some considerations for launching a data science project in public health?

- What is the epidemiological relevance of a project?
- Are predictors and outcomes measurable and interpretable?
- There is no perfect prediction algorithm – what is the trade-off between sensitivity and specificity?
- What are relevant regulations, e.g., on the collection and use of information?
- How is the data quality? The validity of a data science project relies on the quality of data – need to understand how data is captured and to evaluate if there is a data issue (e.g., data not missing at random, data formatting, data representatives, etc.)
- Is the machine learning algorithm under consideration better than a traditional statistical model? Consider quality checks needed before implementing a machine learning model.
- What equity issues might result from the implementation of a machine learning model known to perform better in certain sub-groups than the others?

## What may be limitations of a data science project in public health?

- Predictors may not be connected to underlying causes, and could result in public health practitioners paying less attention and resources to address the root causes of an issue.
- When there is a change in data used to develop prediction modeling, it can be hard to know how this change shifts the prediction pathways and impacts the prediction outcomes. (E.g., in a Chicago project, how an alcohol licence was obtained by a food establishment is used to predict the likelihood of food safety violations. If the mechanism for obtaining licence changed, then this could impact the prediction.)
- It may be unclear where the responsibility lies if an algorithm gives a wrong prediction, resulting in, e.g., change of life quality of patients.

## What can we do to facilitate more data science projects in public health?

- Facilitate access to metadata to help understand what data is available and how they can be linked to create an integrated database

- Shift from paper-based records to electronic data to increase efficiency in data management
- Connect content and technical experts to facilitate more collaborations
- Increase data literacy through public education and communications around the benefits of using aggregated data in public health
- Provide training on data science methods
- Visualize data to help interpret the results and highlight the public health impacts

## More thoughts and questions around data science and public health…

- What worked well in Schenk's experience with implementing data science projects in Chicago* is to know exactly what people want and find a motivated group of people to implement ideas.
- Schenk's experience with data science projects in Chicago also highlights the benefits and need for more open science in public health, including open data and open code.
- Many data science projects that have been implemented successfully, such as the examples Schenk presented, are developed at the municipal level. How can we scale-up such projects from the municipal to the provincial level?
- How can we use data science techniques to inform social determinants of health? Measuring long-term outcomes might not be the best use of machine learning predictive analytics.
- How do we know when we have to collect more data versus try out an alternative analysis?
- Machine learning can give great results with a single dataset, but results may be worse when applied to other data or contexts. There are great expectations, but lack of transparency with models means people can easily make mistakes (e.g., putting outcomes into predictors).
- There is a lot of excitement around applying machine learning, artificial intelligence and data science in public health. Where are their actual limits for producing scientifically sound and ethically viable results?

*In the Feb 5th Grand Rounds presentation of Smart Cities and Public Health, Dr. Schenk showcased several Chicago projects which employ data science methods to predict, optimize, and evaluate public health concerns such as restaurant food safety violations, lead poisoning and West Nile virus outbreaks.