Big data analytics
How it can be used to improve environmental public health outcomes in Canada

Shawna Bourne, Tarun Rihal

ABSTRACT
Changes in overall health and longevity have been realized over time with the introduction of Environmental Public Health (EPH) interventions such as safe food and water supplies, the treatment of waste, and the establishment of standards that have reduced acute illnesses in the population. Over the years, programs and services in Canada gave rise to vast quantities of data that can be harnessed to guide decision-making for continued improvements in EPH outcomes on a population basis as it relates to chronic disease management and prevention. Novel approaches to big data analysis in EPH in the new millennia have the potential to have a similar impact on quality and length of life in the modern context by analyzing the factors leading to chronic illness in the population and improving outcomes. By using big data and machine learning, we can better understand the environmental factors affecting population health. This article presents an opportunity to utilize pre-existing data to explore a new way of assessing the impact of known and unknown health hazards. This article demonstrates how drinking water test results can be used with different big data analytics techniques of increasing complexity to understand and control risk. This approach is beginning to be used to collect new and better organized data with the intent of improving outcomes but can be taken much further to improve population health outcomes for Canadians.

INTRODUCTION
Public health, and environmental public health (EPH) in particular, has given rise to some of the greatest achievements in healthcare history.1-3 These advancements include a reduction in the spread of infectious diseases through improved sanitation and hygiene, decreases in foodborne illness, and gains in maternal and child health. All the above examples of EPH interventions, employed throughout the last two centuries, led to an overall increase in lifespan and quality of life at the population level. These trends have been associated with the eradication of diseases through improvements that target acute illnesses (eg cholera).4 These population health outcomes can be seen in the overall decrease in number of deaths amongst children and seniors, who have the greatest susceptibility to infection when EPH measures are lacking.5

Canadians have employed many EPH interventions that have produced positive impacts such as improved water and sanitation. Over time, Canada’s infant mortality rate has dropped to a record low of 4.5 per 1,000 in 2017 and an average life expectancy of 83 years for females and 78.3 years for males in 2012.6,7 These gains have come as a result of the use of interventions that target a single etiological entity for the infectious process.8 For example, intervening and destroying a pathogen before it can infect a host by using disinfectants in a daycare or hospital and if a host is infected, using antibiotics to kill the agent within the body. Interventions for singular agents of disease have been optimized to minimize the spread of disease in most Canadian communities.

However, today’s top health issues are far more complex. Statistics Canada reports that the top five causes of death in Canada, in order of incidence, are cancer, heart disease, accidents, stroke, and chronic respiratory diseases.9 Aside from reducing mortality, these conditions can also decrease quality of life and add undue cost to the healthcare system with the majority of dollars being spent on hospitals, drugs and physicians.9-11 Each of these conditions have complex causes and do not have a single causative agent; they arise from a complex web of factors, such as genetics, food sources, workplace exposures, and water sources. To continue to improve health outcomes as we move forward, we need new ways of understanding these modern health risks and their interactions in all of their complexity.

BIG DATA AND ANALYTICS POTENTIAL: ENVIRONMENTAL PUBLIC HEALTH
Big data exists in EPH and can be used to inform decisions at both policy and practice levels. Policy makers and practitioners must harness, transform, and analyze big data for it to be useful.12 They must also take into consideration the importance of privacy, confidentiality and transparency through disclosure when employing big data for healthcare decision-making.

The following section explains how big data can be used in EPH. Using the results of drinking water analysis in Canada, this article demonstrates how big data analytics can be employed in the EPH decision-making to advance both knowledge generation and environmental public health practice. Big data analytics range from basic techniques such as trend analysis to complex techniques such as unsupervised machine learning. Unsupervised machine learning uses algorithms to discover patterns in the data through clustering or association, without human input or correction.13 Both techniques can produce new insights based on existing knowledge or by revealing previously unidentified relationships, thereby targeting more specific environmental public health interventions (see Figure 1).
A CANADIAN DRINKING WATER EXAMPLE

In EPH, data is often collected for the purposes of monitoring, risk assessment and regulatory compliance. Over time, vast quantities of data are accumulated to fulfill the monitoring and reporting requirements of legislation (e.g., the Ontario Safe Drinking Water Act). Across Canada, drinking water quality testing occurs on a continuous, daily, weekly, monthly, and/or quarterly cycle. This testing gives rise to reams of structured scientific data that can be mined for decision-making insight.

The most basic analytic tool that can be employed is longitudinal analysis. These trends may include, for example, whether high risk chemical compounds are appearing in water supplies in greater concentrations, or whether water treatment sites are being maintained at a level sufficient to ensure safe drinking water. EPH research analysts could then use this information to identify communities at increasing risk of unsafe water conditions, and in some cases, use advanced machine learning techniques to predict public advisories of unsanitary water conditions in deteriorating water supplies before it becomes a health risk, preventing community outbreaks.

Spatial analytics, or the display of data in reference to location and time, using big data is another area that merits further exploration. There is work already being done to identify “hot spots” using spatial relationships between drinking water quality risk factors. However, this analytical option is limited by the computational power needed for spatial analytics. New big data technologies such as distributed analytics open the door to the development of far more complex geospatial analytic models, such as those that incorporate machine learning techniques, combining data, space and trends over time and predictions about the future. This knowledge could be translated to policy and practice decisions about the ongoing use of particular water sources or treatment options.

Beyond this, big data analytics provide an interesting opportunity to understand health risks that may have already occurred as a result of these chemicals in various drinking water supplies. To explore this, an EPH research analyst may combine data for the same period from the healthcare system and track cancer cases, related deaths and recoveries, and look for correlations in health outcomes. This type of diagnostic analytics could be combined with geospatial information to identify locations where historical incidences of specific types of cancers can be correlated with drinking water sample results. This can provide information for decision-makers about risks inherent in different drinking water supplies.

Big data tools are also more capable of finding underlying patterns in very large and complex data sets than traditional analysis. This subgroup of tools is known as “unsupervised learning”, where machine learning algorithms can find subtle, but important connections in seemingly unconnected data. Big data analytics provides the opportunity to study and understand the web of factors and interactions that affect the health outcomes in Canadians without the limitations of the Gold Standard Methods (i.e., controlled experiments) that are useful for lab studies but not as useful for understanding how chemical interactions can impact outcomes in the real world. Big data analytics and machine learning can help to solve this problem by analyzing the impacts of long-term, low level exposures in vivo.

In addition to the parameters identified previously, including data on the incidence of cancer cases, relevant geographical information and drinking water test results, the data being analyzed could be further supplemented with additional data points that already exist in Canadian data repositories. This stored data includes information on chemical exposure from pollution sources like air quality index readings by time periods and locations, and community demographic data such as age and sex of patients. This attribute data can be used to indicate specific vulnerabilities and exposure risks in the cancer cases being assessed, while using specific approaches to protect privacy when dealing with large volumes of confidential data. Incorporating privacy best practices where interrelationships exist between disease and personal variables will help to bring forward weak or unknown correlations that would otherwise remain hidden in the data.

With more public institutions moving towards “open data” principles and methods, the availability of this data is at an all-time high. New insights could potentially be drawn about any of the top five causes of death and lost quality of life in Canada and with this new information, EPH policy and practitioners could be successful at keeping Canadians healthy at lower cost and with improved quality of life.

GETTING TO BIG DATA ANALYTICS IN ENVIRONMENTAL PUBLIC HEALTH

Big data analytics require large-scale volumes of data from single or multiple sources, the right competencies and adequate infrastructure. Currently, there are large volumes of data being collected that could add to increased EPH knowledge in Canada. Much of this data relates to factors known to influence the top five diseases affecting health outcomes in Canada, including cancer.
There are researchers currently engaged in interesting big data and machine learning analytics research to identify high risk food establishments and to uncover the factors contributing to poor water quality at small drinking water systems on First Nations water supplies in Canada. There are many other interesting research initiatives in progress.

CONCLUSION

There is data available that may be the key to understanding the health effects of ongoing low-level exposures to contaminants in Canadian communities. These data points exist in different databases and are owned and operated by different levels of government institutions, and private industry. This information is currently unconnected and unexplored for valid reasons including privacy, intellectual property rights and other legal and moral concerns. Although challenging to overcome, these concerns are not insurmountable and can be addressed with proper planning and consideration. Next steps include governments, non-government organizations and industry to explore this data, and uncover new truths about exposures and health outcomes. Getting to big data analytics in EPH will support public and population health goals, drive down costs and ensure that our healthcare actions and dollars are employed in the best possible way.

REFERENCES


