# Machine learning uncovers missing information about ethnicity and Aboriginal status in population health data: study

November 18 2020

Kai On Wong found that machine learning can be used to predict ethnic background from public health data, which would help fill an information gap and could eventually inform policies aimed at reducing health and social inequities. Credit: University of Alberta

Machine learning can be used to fill a significant gap in Canadian public health data related to ethnicity and Aboriginal status, according to

research published today in *PLOS ONE* by a University of Alberta research epidemiologist.

Kai On Wong, senior data scientist at the Real World Evidence unit of the Northern Alberta Clinical Trials and Research Centre (NACTRC), said ethnicity and Aboriginal status are recognized as key social determinants of health but are often not reported in large databases that track acute and chronic diseases such as asthma, influenza, cancer, cardiovascular diseases, diabetes, disability and mental illness.

"If a database currently lacks ethnicity information, we will not be able to tell whether certain ethnic groups have higher rates of disease or worse clinical outcomes," Wong said, "This is a way to unlock that missing dimension from existing data sources, which may help us understand, monitor and address issues such as social inequities and racism in Canada."

Wong created a machine learning framework to analyze the names and geographic locations of 4.8 million people surveyed in the 1901 census, examining features such as spelling and phonetics to predict whether they belonged to one of 13 ethnic groups.

"Different ethnic and linguistic groups have different manifestations of features such as how the name sounds, how many letters in the name, how many vowels and unique letter sequences, and so on," said Wong, who created the program and shared it as a public GitHub repository as part of his doctoral thesis at the U of A's School of Public Health.

"Machine learning is like having a team of agents who are given vast amounts of information. They are instructed to detect and retain useful patterns to solve practical problems such as predicting the ethnicity from the readily available information," he said.

Wong said the program performed best at identifying individuals of Chinese, French, Japanese and Russian heritage based on name only, while the accuracy was improved for the Aboriginal classification when locations were also included.

Both the World Health Organization and the Government of Canada recognize ethnicity and Indigeneity as determinants of health, along with other factors such as income, education and gender. Wong first became interested in inequities in health care that affect Indigenous groups when he served as acting territorial epidemiologist for the Government of the Northwest Territories.

Wong said while American health records tend to include questions about ethnicity, this information is not collected consistently in Canadian databases ranging from hospital discharge records to cancer registries.

By using machine learning to uncover this missing information, researchers and policy-makers will be able to learn more from existing records rather than having to carry out new population-level surveys, which are expensive and time-consuming.

"A future step forward will be to validate this research with real-world applications using health evidence augmented with ethnicity generated by the machine learning framework and comparing it with existing literature, particularly on health and social inequities," Wong said.

Wong recommends first updating the ethnicity prediction tool using more recent census information and testing its accuracy when applied to various health records.

"It is unrealistic to expect machine learning predictions to be 100 percent accurate at all times," Wong said. "The goal is to make predictions that are accurate and generalizable enough to discern

underlying patterns in a meaningful way for a particular problem or application."

**More information:** Kai On Wong et al, A machine learning approach to predict ethnicity using personal name and census location in Canada, *PLOS ONE* (2020). DOI: 10.1371/journal.pone.0241239