Public Health Applications Using Big Data and Machine Learning Methods:

Name- and Location-based Aboriginal Ethnicity Classification

and

Sentiment Analysis of Breast Cancer Screening in the United States Using Twitter

by

Kai On Wong

A thesis submitted in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

EPIDEMIOLOGY

School of Public Health

University of Alberta

ABSTRACT

Applications using big data and machine learning techniques are transforming how people live in the 21st century, however they are generally underutilized in public health compared to other domains. We proposed and conducted two independent studies to investigate how big data and machine learning techniques may serve important functions to address different public health challenges in North America.

In *Name- and Location-based Aboriginal Ethnicity Classification*, we developed and tested the classification performance of a machine learning method to predict individuals' Aboriginal status using name and location information from the 1901 Canadian census. Our automated approach has yielded good classification results, especially for a number of Aboriginal (all-inclusive) and sub-Aboriginal (such as First Nations, Algonquian, and Kootenay) statuses. The classification performance for predicting ethnicity status of these four Aboriginal groupings ranged between 0.99-1.00 in accuracy, 0.99-1.00 in ROC, 0.63-0.65 in sensitivity, 0.99-1.00 in specificity, 0.78-0.86 in PPV, and 0.99-1.00 in NPV in the validation sets. The demonstrated application illustrated that using high decision boundary values resulted in predicted First Nations-specific prevalence statistics closely approximated to the true underlying prevalence.

In *Sentiment Analysis of Breast Cancer Screening in the United States Using Twitter*, we slightly modified the existing VADER sentiment classifier to automatically classify the sentiment of breast cancer screening-related tweets into neutral, positive, and negative. Extensive data visualization was conducted to illustrate the temporal (via time-series plot), geospatial (via point, hot spot, and quintile maps), and thematic (via word-clouds) patterns of breast cancer screening sentiment in the U.S. The ecological associations between the averaged sentiment

scores and percentage of breast cancer screening uptake at the state level were examined, and significant inverse relationships ($p<0.05$) were found between negative sentiments and recent uptakes of mammogram and clinical breast exam.

PREFACE

This thesis is an original work by Kai On Wong, MSc., with primary supervision of Dr. Yutaka Yasui and co-supervision of Dr. Faith Davis. Chapter 1 will summarize the background information pertaining to big data and machine learning, as well as their relevance to epidemiologic research. Chapter 2 will lay out the general research directions and study components (including literature review, as well as research rationale, objectives, questions, data sources, and methods) for the two independent studies: *Name- and Location-based Aboriginal Ethnicity Classification* and *Sentiment Analysis of Breast Cancer Screening in the United States Using Twitter*. Each of these studies utilizes different epidemiologic concepts and techniques, as well as big data analytics and machine learning methods in order to address two major public health issues in North America.

Chapter 3 of this thesis will be submitted as Wong KO, Davis FG, Zaïane OR, and Yasui Y. *Position Paper: Concerning Data Inadequacy on Ethnicity and Race in Canada - Description and Recommendations*. I was responsible for constructing the discussion topics, formulating scientific opinions, gathering research evidence, deriving recommendations, and drafting the manuscript. Dr. Yasui oversaw my research and provided iterative manuscript editing and feedback. Dr. Faith Davis and Dr. Osmar Zaïane have edited the manuscript and provided recommendations.

Chapter 4 of this thesis will be submitted as Wong KO, Zaïane OR, Davis FG, and Yasui Y. *Name- and Location-based Aboriginal Ethnicity Classification.* Chapter 5 of this paper has been submitted and published in the Proceedings of the 8th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (2016), as Wong KO, Davis FG, Zaïane OR, and Yasui Y. *Sentiment Analysis of Breast Cancer Screening in the*

United States Using Twitter. For both papers of Chapters 4 and 5, I was responsible for the conceptualization and development of research hypothesis, data collection/retrieval and preprocessing, general-purpose and statistical programming, conducting the analysis, interpretation and presentation of study results, and manuscript drafting and final submission. Dr. Yasui and Dr. Davis are the supervisory authors involved in editing manuscript and providing feedback from the conceptualization of research topics to manuscript submission. Dr. Zaïane provided machine learning-domain specific knowledge regarding the proposed study designs, as well as feedback on the manuscripts.

DEDICATION

To my parents, Hin and Tin,

my sister Christy and her family,

and my wife Jacqueline.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude and appreciation to many people who have, directly and indirectly, supported my journey in this doctoral program. First, I am grateful for the opportunities to work with and learn from my primary supervisor, Dr. Yutaka Yasui. He has been tremendously supportive throughout the years. He has provided me the research and intellectual autonomy to freely pursue a unique research path in epidemiology that inspired my core interests. He has continuously provided constructive feedback and recommendations at various research stages. He has led by countless examples of what it means to be a truly remarkable scientist by developing thorough knowledge of abundant research tools and always upholding best scientific practices as guiding principles. His passion and achievements in research, mentoring new researchers, and making positive differences in scientific communities and the general public are phenomenal and have set high standards for his students and team members, including myself, to constantly look up to.

I would like to sincerely thank Dr. Faith Davis for being my co-supervisor. She has made her support available in various ways and provided important suggestions and feedback in my research development and manuscript writing. She has given me a lot of intellectual freedom and opportunities to survey potential research topics, and has provided valuable feedback in my projects through her perspective and expertise in public health and cancer epidemiology. I would also like to thank Dr. Osmar Zaïane for being my thesis committee member who has shared his extensive expertise in the domains of computer science and machine learning. He has given important insights, confirmation, and recommendations of machine learning-related concepts and techniques there were invaluable for my research progress.

I would also like to give my appreciation to the members of Dr. Yasui's research team whom I have fond memories from sharing ideas, experiences, and friendships throughout the years. I am delighted to have met many fantastic professors, staff members, and students who are the essential catalysts driving the successes of the School of Public Health. Special thanks to Dr. Kue Young, Dr. Ambikaipakan Senthilselvan, Dr. Yan Yuan, Dr. Marcy Winget, He Gao, and Sohaib Mohammad for their professional and moral supports.

I like to express my gratitude to my parents, Hin and Tin, and my sister, Christy, for their encouragement. They have always been patient and supportive of my education dreams and career choices. I am truly blessed to have my wife, Jacqueline, in my life. She has always been there for me during my ups and downs. Her unconditional support, understanding, and love have led me to believe we can overcome any challenges and achieve any life goals together.

I would like to gratefully acknowledge my funding supports that I have received from a number of organizations over the years including the CIHR Doctoral Research Award from Government of Canada, President's Doctoral Prize of Distinction and Queen Elizabeth II Graduate (Doctoral) Scholarships from University of Alberta, as well as research funding from AMII (then AICML) and research assistant salaries from University of Alberta and Alberta Health Services.

Table of Contents

List of Tables

List of Figures

List of Supplementary Tables

List of Supplementary Figures

List of Abbreviations

ACOG	American College of Obstetricians and Gynecologists
ACS	Aboriginal Children's Survey
AI	Artificial intelligence
ANEW	Affective Norms for English Words
API	Application Programming Interface
APS	Aboriginal Peoples Survey
ASCO	American Society of Clinical Oncology
BCE	Before Common Era
BDA	Big data analytics
BRCA	BReast CAncer susceptibility gene
BRFSS	Behavioral Risk Factor Surveillance System
CBE	Clinical breast exam
CCHS	Canadian Community Health Survey
CCR	Canadian Cancer Registry
CDC	Centre of Disease Control and Prevention
CEO	Chief executive officers
CES	Centre for Education Statistics
CHI	Canada Health Infoway
CHMS	Canadian Health Measures Survey
CI	Confidence intervals
CIHI	Social determinants of health
CMA	Canadian Medical Association
CMEC	Council of Ministers of Education-Canada
CORR	Canadian Organ Replacement Register
CPERS	Canadian Patient Experiences Reporting System
DAD	Discharge Abstract Database
DT	Decision trees (classifier)
EDS	Ethnic Diversity Survey
Edu	Education
EMR	Electronic medical records
FN	First Nations
fn	False negative
fp	False positive
GenHlth	General health
GI	General Inquirer
GPS	Global positioning system
H1N1	Hemagglutinin Type 1 and Neuraminidase Type 1
HBM	Health belief model

HIR	Health Insurance Registries
IBM	International Business Machines
IDC	International Data Corporation
IEEE	Proceedings of the Institute of Electrical and Electronics Engineers
IG	Information gain
IoT	Internet of Things
IT	Information technology
KNN	K-nearest neighbours
LIWC	Linguistic Inquiry and Word Count
LR	Regularized logistic regression (classifier)
Mamm	Mammogram
ML	Machine learning
MRI	Magnetic resonance imaging
NLSCY	National Longitudinal Survey of Children and Youth
NoSQL	Not Only Structured Query Language
NPDB	National Physician Database
NPHS	National Population Health Survey
NPV	Negative predictive value
PHAC	Public Health Agency of Canada
PPV	Positive predictive value
Q1, Q2, …	Question 1, question 2, …
REST	Representational State Transfer
ROC	(Area under the) Receiver-Operating Characteristics curve
ROI	Return on investment
RT	Retweet tag
S1, S2, …	Study 1, study 2, …
SCN	SenticNet
SentCom	Composite VADER sentiment score
SentNeg	Negative VADER sentiment score
SentNeu	Neutral VADER sentiment score
SentPos	Positive VADER sentiment score
SES	Socioeconomic status
SQ1, SQ2, …	Supplementary question 1, Supplementary question 2, …
SVM	Support vector machines (classifier)
SWN	SentiWordNet
tn	True negative
tp	True positive
U.S.	United States
URL	Uniform Resource Location
USPSTF	United States Preventive Services Task Force
VADER	Valence Aware Dictionary for sEntiment Reasoning
WHO	World Health Organization
WSD	Word-Sense Disambiguation

Chapter 1 – Introduction

1.1 Big Data

1.1.1 Background

Modern lives in the 21st century are immersed by the flood of digital data. Humans have long been using data to decipher patterns and obtain useful knowledge about the physical and social spheres (1). One of the earliest human practices of data storage was dated in 18,000 Before Common Era (BCE) when Palaeolithic tribespeople carved notches into sticks and bones to keep track of trading activities and surplus supplies (2). While the rate of data production has grown steadily throughout human history, recent decades have vastly outpaced the past. The term "information explosion" refers to the rapid increase in the amount of published data and information in recent decades (3). Although the term "information explosion" was first used in the 1940's, the scope of its impacts on human lives has not been realized until the 1980's, which marked the beginning of rapid technological advancement in computer software and hardware (4). In recent decades, portable electronic devices (such as laptops and smart devices) are capable of capturing, storing, and distributing large amount of information and these devices are made affordable and accessible to a massive user base (5). As a result, an unprecedented amount of digital data is produced, consumed, and reproduced constantly.

The total volume of data in the world doubles every 20 months (6). Between the dawn of human civilization and year 2003, a total of five exabytes of data was created (note: one exabyte is equal to one billion gigatbytes), yet the same amount of data was produced in only two days in 2010 (7). In 2013, it was estimated that 90% of all existing data in the world was created within the previous two years (8). The International Data Corporation (IDC) Digital Universe Study

predicts that between 2009 and 2020, digital data will grow by 44-fold to 35 zettabytes per year (note: one zettabyte is equal to one trillion gigabytes) (9). The increasing ease in obtaining personal computing devices and connecting to the World Wide Web are the major driving forces behind the digital data surge. By the end of 2015, there were more than seven billion mobile cellular subscriptions in the world, corresponding to a 97% global penetration rate (10). Global mobile broadband penetration has reached a historical high at 47% in 2015. The proportion of households with Internet access across the world has increased from 18% to 46% from 2005 to 2015 (10).

In addition to the typical web surfing-related uses of Internet, the Internet of Things (IoT) and cloud computing have also contributed to a large amount of digital footprints. The IoT refers to the interconnectedness of physical (wearable or portable) devices, vehicles, homes and buildings, and other items that are embedded with electronics, software, sensors, actuators, and network connectivity that enable these objects to collect and exchange data (11). Cloud computing refers to a type of Internet-based, on-demand computing enabling network access to a shared pool of configurable computing resources (i.e., networking, servers, storage, data center space, applications, and other information technology (IT) resources) with other remote devices (12). Furthermore, many commercial and non-commercial organizations have also employed partial or full digitization in workflows by collecting, storing, and retrieving data electronically (13). Today, constant streams of data are produced from almost all facets of life, including electronic medical records (EMRs), business transactions, weather sensors, space imaging, online social media contents, digital pictures and videos, cell phones' global positioning system (GPS) and other signals, as well as countless other examples. Modern lives are often referred to

as living in the "big data" era due to the encompassing, ongoing massive data production and consumption that transpire virtually all aspects of life (14).

1.1.2 Definitions of big data

The term "big data" has gained a lot of attention in the past decade (15). It has been increasingly used by the media and professionals in technical communication to convey the concepts of very large datasets in areas including healthcare, biology, surveillance, finance, commercial industries, online web, and social and news media (9). However, there is not a universally-accepted definition of big data, resulting in the term "big data" being loosely used, misused, and overused (16, 17). It may refer to a single large dataset, or it may refer to the conceptual totality of existing data. It may refer to large dataset with complex underlying relationships, or it may refer to analysis using techniques (such as predictive analysis) commonly-used to analyze large datasets.

One of the first articles that used the term "big data" is *Application-controlled demand paging for out-of-core visualization*, by Cox and Ellsworth, published in the Proceedings of the Institute of Electrical and Electronics Engineers (IEEE) 8th conference on Visualization in 1997. In this article, "big data" is used to describe datasets that "do not fit in main memory (in core), or when they do not fit even on local disk, the most common solution is to acquire more resources" (15). In other words, this first use of "big data" was confined to two quantities, namely processing/storage capacity of computer and amount of data, and their relative relationship. A given dataset of a fixed size may be "big data" for one computer but not for another. In following decades, computers' processing and storage power increased exponentially, pairing with advances such as out-of-core and parallel computing, and also larger amount of data able to be processed and stored at once. Though at the same time, this very phenomenon of soaring

computing power also led to more and more data produced as primary- or by-products. Beyond the growth in data size, new data types and formats have also emerged, including books, journal articles, metadata, health records, audio, videos, analog data, images, work documents, environmental sensors, and unstructured text such as emails, text messages, web pages, word-processor documents, and online social media and forum postings (17, 18). These are primarily examples of unstructured data, which are data without a predefined data model or predefined organizational/relational structure. Currently, there is a lack of standardized scheme of how unstructured data should be extracted, stored, linked, organized, retrieved, processed, and analyzed. In addition to the characteristics of big data per se, issues such as limitations of current data processing systems and finding the appropriate analysis models and methods present unfamiliar challenges to data managers, architects, and analysts. Jin *et al.* (2015) divided the complexity of handling big data into three-fold (19):

Data complexity

- common characteristics of big data including diversified types and patterns, complicated inter-relationships, a large set of features, and varying data quality;
- laws of underlying data distribution, representativeness of population segments, heterogeneity of effects (interactions), and relevant assumptions (i.e., linearity and independence), uncertainty, and biases may be unclear, violated, or infeasible to be assessed;
- aforementioned challenges lead to difficulties in perceiving, representing, understanding, processing, and describing different aspects of big data.

Computational/analytic complexity

- common characteristics of big data including multi-sources, high volume, and fast-changing often render traditional computing methods (such as parametric statistics) inappropriate;
- new computing methods need to be able to examine data without certain statistical assumptions such as independent and identical distribution (i.d.d.) and adequate (i.e., random) sampling of data;
- large set of features, feature extraction, and ways of modeling features lead to very large dimensional spaces (relative to the obtainable sample size), frequently named as the curse of dimensionality, which may exceed processing power and/or render computing time unacceptable;
- new methods required to address analytic issues such as insufficient samples, uncertain data relationships, and unbalanced distribution of value density;
- new methods required to reduce the size of big data to be minimally-large enough to provide meaningful insights on demand;
- overall, new approaches for big data computing need to develop novel and highly-efficient computing paradigms which require algorithms for distributed and streaming computing and big data-oriented framework where communication, storage, and computing are well integrated, scaled, and optimized.

System complexity

- processing systems and supporting infrastructure of big data needed to handle large data volume and diversified data types and applications with competent and sufficient team members, physical space, software, and hardware;

- computing systems need to be able to perform accurately and reliably under large processing loads over prolonged duty cycles and oftentimes with real-time requirements;
- operational inefficiency and energy consumption in view of cost and environmental impacts.

Understanding these complexities of big data is important regarding its definition since the conceptual boundary of big data has evolved not only to reflect the large data volume, but also the high levels of complexity in data characteristics, as well as the processes in management, retrieval, and analysis. In recent years, the meaning of big data can be roughly grouped into two definitions. A narrower definition refers to databases that are too large and its relationships too complex to be handled by conventional structured relational databases (20). A relational database organizes data into one or more tables (called "relations") in the form of columns and rows, generally with a unique key for each row. The relationships between database tables are well-defined in relational databases, allowing for communication and sharing of information between tables. A relational database model uses column names as metadata to characterise a given dataset (21). In the case of big data deriving by different data sources, it presents some challenges such as non-standardization of metadata (i.e., labeling as "gender" in one source while "sex" in another) and automatic mapping of metadata (i.e., mapping "patient name" in one source to "first name" + "last name" in another) (21). Thus, the narrower definition of big data frequently includes a discussion of Hadoop, MapReduce, NoSQL (or Not Only Structured Query Language) or similar data management technologies that are used to solve the aforementioned challenges of handling with large volumes of unstructured data (20). The wider, more commonly-used, definition of big data does not require the "unstructuredness" of data or

the use of non-relational database technologies (20). It focuses on the large amount of data, in terms of quantity, timeliness, variety, and application possibility that may provide value to data holders and users.

1.1.3 The five V's of big data

The wider definition is in alignment with the key characteristics of big data introduced by the International Business Machines (IBM) as the five V's (22, 23):

Volume – refers to the large quantity (or volume) of data, although there is no consensus of any specific cut-off size. Large volume refers to a large number of entities (number of rows) and it could have a large number of variables (number of columns).

Velocity – refers to the fast and sometimes varying speed of data generation. Many data sources such as social media, credit card transaction, and certain tracking mobile apps are creating and storing data in continuous and real-time fashion.

Variety – refers to the myriad of data types, formats, and sources existing and available for use. Unlike in the past, the large majority of the world's data now is unstructured and cannot be easily stored in tables or relational databases (i.e., photos, videos, and social media updates).

Veracity – refers to the degree of trustworthiness of data. Due to the uncertainty of data obtained from data sources with unknown or unclear sampling or data collection methods, the underlying quality of the data may be compromised or unknown.

Value – refers to the potentials to obtain valuable information and knowledge that could inform decision-making processes. Big data applications often leverage multiple, independent data sources together or invent innovative uses of existing data to extract previously unknown information/knowledge that may be valuable to stakeholders.

Although the five V's is only meant to be an imperfect and subjective guide, they have been frequently used to qualitatively classify whether a dataset and its applications should be labeled as "big data" (8). In general, the more number of V's a dataset possesses, the more confident one can consider it as big data. On the other hand, Bernard Marr (2015) argued that of the five V's, value is the most important aspect of big data (23). The ultimate value of big data applications will be determined by one or more of three guiding questions (9): 1) Does it provide more useful information? 2) Does it improve the fidelity of existing information? 3) Does it improve the timeliness of actionable responses or decision-making processes? This dissertation does not attempt to redefine what big data entails, but to convey the important concept that even though the definition of big data is imperfect, the impact of big data are tangible and far-reaching in many health sectors including healthcare, health economics, outcome research, and public health and epidemiology (24-27).

1.1.4 Machine learning and big data

Big data analytics (BDA) is the process of collecting, organizing, and analyzing big data to extract valuable information and discover new knowledge (21). One of the main objectives in conducting BDA is to make predictions of future outcomes, also called predictive analytics. When the predicted outcomes are on a continuous numerical scale, it is called a *regression* task; when the predicted outcomes fall into discrete categories/groupings, it is called a *classification* task. Traditional approaches with regression-based models (i.e., linear regression) for BDA are limited due to the frequently-encountered big data complexities such as violation of linear and independence assumptions, heterogeneity of effects, and curse of dimensionality (28). Traditional regression models may be modified or entirely forgone in favour of a group of more

flexible modeling techniques for predictive analyses belonging to the methodological approach called machine learning (ML).

Machine learning is a type of artificial intelligence (AI) that provides computers the "ability to learn without being explicitly programmed" (29). This fits well with BDA as designing and programming explicit learning algorithms are usually impractical and impossible due to the complexity of big data. Instead, ML algorithms preform an automated search, either stochastic or deterministic, to obtain the optimal model (or best fit) within a defined feature space (29). The outcome of a deterministic model is fully determined by parameter values and initial conditions, whereas stochastic models maintain some inherent randomness. Contrasting with the traditional regression models which rely on pre-specification of a model structure and assumptions, the search performed in ML is primarily data-driven (6). Generally, ML algorithms can be classified into two categories: supervised (or predictive) and unsupervised (or descriptive) (30). In supervised learning, each sample has a set of feature values and a "correct" (or true) outcome value (i.e., number or class label) (31). In contrast, such "correct" results are non-existing or not provided in unsupervised learning, instead its ML algorithms explore the inherent data structure by grouping samples into similar categories (or clusters) based on similarity of input values. The general working mechanism and components of supervised learning is illustrated in **Figure 1-1**, extracted from Talwar and Kumar (2013) (32).

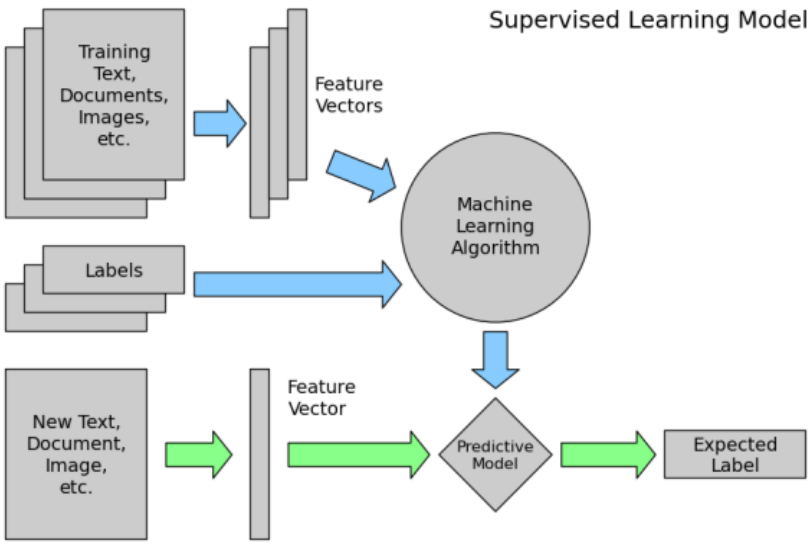


Figure 1-1: Block-diagram showing the components and working mechanism of supervised learning (Extracted from (32)).

1.1.5 Overfitting and hold-out validation

Discussions on supervised learning below will focus on binary classification since the research studies of this thesis are examples of classification, not regression, tasks. There are many supervised ML algorithms, including regularized logistic regression (LR), support vector machine (SVM), decision trees (DT), naïve Bayes (NB), and k-nearest neighbours (KNN) (6). The performance of specific ML algorithms may differ in different data and analytic environments. Influencing factors include size of training sample, size of feature set, linear separability of data, existence of highly correlated features, desired computation speed, available processing memory, ease of model interpretability, and acceptable level of model variance (33-35).

In supervised learning, a phenomenon called overfitting may occur when a ML algorithm is allowed to fit the model to a given data too excessively (36). Overfitted model maps the features too closely to the true output to a degree that even random noise component of the

output (which is not predictable) is captured for the specific set of samples at hand. Capturing the random noise component in the specific set of sample does not help predicting output of a new data sample. Such model will appear to have high predictive performance within the training sample, but its predictive performance in a new data sample will suffer from the lack of generalizability since the modelled random noise does not exist in new data. Overfitting typically occurs when the size of the training sample is too small relative to the specified model complexity, or when the ratio $\frac{Model\ complexity}{Training\ sample\ size}$ is too high. Examples of how this can take place include 1) high noise level in training data, 2) incremental refinements of model over time with increasing data inputs, and 3) lack of proper stopping rules in learning (such as allowing decision trees to grow to full depth) (37, 38). A number of methodological strategies and techniques is generally implemented to minimize potential overfitting such as collecting and using more data, reducing the number of features, applying regularization, parameter-tuning during cross validation, and if applicable, pooling multiple ML results using ensemble methods (6, 39-42).

Cross validation is a model evaluation technique that can assess the generalizability of trained machine learning models within a given dataset to a new data sample. It is often used in ML to minimize and assess potential overfitting since cross validation separate the full data into independent subsets (i.e., training and validation sets). There are different variants of cross validation techniques including 1) holdout validation, 2) *k*-fold cross validation, and 3) leave-one-out cross validation (43). Let a labeled dataset to contain the full set of input-output pairs ($\boldsymbol{x}$, y), which specifies y to be the ground-truth output for feature vector $\boldsymbol{x}$. In holdout validation, the full dataset is split into two: a training set and a validation (or testing) set (6). Completely independent from the validation set, machine learning takes place in the training set by searching

for the optimal parameter values to best map features $\boldsymbol{x}$ to the label y via an optimization objective (i.e., minimization of loss function or maximization of information gain). The k-fold cross validation divides the dataset into k subsets, and the holdout method is carried out repeatedly for k times (43). Unlike the holdout method, k-fold validation gives every data point a chance to be trained by the ML algorithm. The leave-one-out method is a logical extension of the k-fold method. Leave-one-out cross validation iteratively trains all data points except for one, and it makes a prediction for that data point.

1.1.6 Bias-variance trade-off

The key mechanism behind ML is to minimize prediction error/misclassification without overfitting. The total expected prediction error can be expressed as $Prediction\ error = Bias^2 + Variance + Irreducible\ error$ (44). Bias results from erroneous assumptions or model specification in the training algorithm. A model with high bias may miss important relationships between the features and output, resulting in underfitting. Variance results from sensitivity to slight fluctuations in the training sample. A model with high variance may suffer from overfitting via modeling the random noise from training sample. Irreducible error is the inherent noise component in the distribution of outputs y that can not be reduced via model or sampling specification (44). In supervised learning, the objective is to train a model to not only correctly capture important patterns, but also generalize to new data. In the most ideal scenario where a model is perfectly specified and infinite data is available, both bias and variance will be reduced to zero. However, real world analyses consist of imperfect models and finite data, leading to the trade-off between reducing bias and reducing variance (45). For example, models with low bias will capture relationships in the training sample vigorously. Though these models are generally more complex (i.e., more features) which may lead to high variance, and thus lower

generalizability in new sample. In contrast, models with low variance are often simpler models, which may result in high bias.

1.1.7 Predictive performance metrics

The predictive performance (and degree of prediction error/misclassification) can be summarized by different performance metrics such as accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and area under the Receiver-Operating Characteristics (ROC) curve. Each of these metrics has its own strengths and limitations. Some of these metrics should be considered as more important than others depending on the goal and nature of the classification task and characteristics of existing data. For example, accuracy ($Accuracy = \frac{Number\ of\ correct\ classification}{Number\ of\ total\ classification}$) assumes equal cost for misclassification of target class and misclassification of non-target class. In the case of imbalanced data, data with high over-representation of one class over the other class, accuracy is a poor indicator to evaluate models' predictive performance. This is because a ML classifier can simply and blindly assign all/most data points to the majority class to attain high accuracy, without much learning. Sensitivity ($Sensitivity = \frac{Number\ of\ correctly\ classified\ target\ sample}{Number\ of\ all\ target\ sample}$) and PPV ($PPV = \frac{Number\ of\ correctly\ classified\ target\ sample}{Number\ of\ all\ people\ classfieid\ as\ target\ sample}$) would serve as better predictive performance indicators than accuracy in unbalanced data. Sensitivity focuses only on individuals from a certain target class (i.e., minority class), and computes the proportion of all the target class individuals that are correctly captured by the ML classifiers. PPV focuses only on individuals predicted to belong in certain target class (i.e., minority class), and computes the proportion of all the individuals predicted to be in a target class that actually belong to that class. Further

discussion on the aforementioned predictive performance metrics will be carried out in Chapter 4.

1.2 Big Data and Epidemiology

1.2.1 Relevance of big data for epidemiology

Public health is a mix of sciences and skills focusing on the preservation and improvement of the health of all people through preventive and often group-level, in contrast to individual-specific, interventions (46). Epidemiology is a basic science of public health that studies patterns, causes, and effects of health and disease conditions in populations (47). It contributes to public health by accumulating scientific evidence to gain a better understanding of nature and mechanisms of health and disease conditions, which in turn, may be translated into cost savings in the health system and improved population health outcomes through informed decision making (26). Epidemiologists are interested in describing health patterns and testing hypotheses to identify the underlying roles (such as risk factors, confounder, and effect modifiers) of interested factors within disease causal pathways. To perform these functions, epidemiologists take part in designing and evaluating the validity of research studies, as well as collecting, examining, manipulating, and analyzing large amount of health-related data. Using the five V's big data guide, many health-related datasets possess big data characteristics. Due to the common interest in studying large populations, many epidemiologic research and applications use high volumes of data. For instance, Canadian censuses are generally used to derive the denominators when computing national health statistics (i.e., incidence and prevalence rates). The number of people captured in the Canadian census is in the tens of millions. Regarding data variety, factors contributing to disease causal pathways span from proximal (such as diet, physical activity, and smoking status) to distal (such as socioeconomic status (SES),

ethnic culture, residential location, and health laws and policies) (48). Hence, health data used by epidemiologists often come from various domains, including government, laboratory, clinical, genomic, environmental, interview, survey, surveillance, and administrative. Health data can be qualitative or quantitative, prospective or retrospective, controlled or observational, and primary or secondary (49), thus furthering the variety of health-related data.

Epidemiologists often use secondary data, which are already-existing data to be used for purposes not originally intended. Secondary data have many advantages over primary data such as more timely access and lower economic costs (50). However, the major drawback is that since data is not collected or managed directly, researchers will have to endorse and/or address a greater level of uncertainty in data quality with respect to biases and errors. This suggests a lower degree of veracity in using secondary data sources especially when comprehensive and detailed data profile is unavailable. Non-secondary data sources are not guaranteed to be free of data uncertainty either, as it depends on many factors including how the study is designed and actually carried out. In terms of value, while the tangible benefits derived from any particular single epidemiologic study may not be easily or directly realized, the cumulative efforts by public health professionals and epidemiologists have led to many significant breakthroughs in improving the health of the general population, including vaccination, motor-vehicle safety, occupational safety, control of infectious diseases, cardiovascular disease prevention, cancer prevention, healthier food, maternal and infant health, family planning, fluoridation of drinking water, recognition of tobacco as a health hazard and tobacco control, inclusive health coverage, and public health preparedness and response, all of which contribute to improved life expectancy and quality of life (51). While it has intrinsic value to understand a disease mechanism per se, the more tangible value from epidemiologic endeavors usually derives from working with decision

makers to translate relevant epidemiologic knowledge into actionable plans. At its core, the big data era should not be shockingly new for most epidemiologists because they have been trained and using data of high volume (52-55), great variety (56-58), low veracity (53, 59, 60), and potentially good value (51, 61, 62) to tackle public health issues in the past. However, a number of important aspects of big data are rather novel for public health and epidemiology.

1.2.2 New aspects and opportunities for epidemiology

Despite a degree of familiarity of handling large datasets, the big data era presents new challenges and opportunities for epidemiologists. Presently, health data has even higher volume, greater variety, faster velocity, and more uncertain veracity. An example of the large surge in data volume is demonstrated by Canadian primary care physicians' increased use of EMRs, rising from 23% to 73% between 2006 to 2015 (63). Likewise, screening and diagnostic test results are now being digitalized in many healthcare locations (64). Health and monitoring device and smartphone apps have also drastically expanded the volume of health-related information. A larger data size likely increases statistical power for association studies, thus possibly increasing the likelihood of obtaining statistically significant finding without substantive biological or public health relevance (i.e., small effect size) (25). For variety, in addition to the more traditional aforementioned data types, epidemiologists are now presented with new data types such as streaming (i.e., home monitoring, telehealth, handheld and sensor-based wireless or smart devices) and web/social networking (i.e., search engine usage and online microblogging or messaging sites) which may hold a largely-untapped potential for public health knowledge discovery that was not possible before (65). Furthermore, these databases often collect unstructured data type. For example, up to 80% of clinical data (such as documents, images, clinical or transcribed notes) and almost all of the health-related social media messages

are unstructured (66). The wide variety of data sources will allow epidemiologic studies to examine novel and complex interactions, but at the expense of the needs to effectively select and model complicated variable sets (25).

Among the 5'Vs, the increment in data velocity is arguably the most revolutionary and unfamiliar aspect to most public health researchers (25). Comparing to the real-time and continuous data generation from many big data sources, traditional health data sources generally operate at a much slower and infrequent pace. While streaming data may help design more dynamic and responsive interventional and surveillance applications, the lack of familiarity and formal training in handling a large amount of real-time data may hinder epidemiologists' ability to adequately utilize such data (67). In terms of veracity, emerging big data sources naturally pose many new and unexplored questions to epidemiologists regarding data quality. Namely, many online social media (i.e., Twitter and Facebook) and forum sites have unknown population distribution and representativeness. When data owners allow public access to a subsample of their full data (for example, Twitter releases 1% of all tweets via the Representational State Transfer (or REST) Application Programming Interface (API)), it is uncertain how the subsample is selected. In addition, online social media and forum may contain a considerable portion of duplicated, fake, or abandoned users, all of these may lead to inaccurate study inferences. Overall, big data presents many unprecedented channels for public health and epidemiologic innovation and experimentation which should be proceeded with reasonable skepticism and adherence to solid methodologic foundation.

1.2.3 Big data challenges in epidemiology

In recent years, there is a number of notable examples where big data has been incorporated in health-related applications (66), including disease outbreak surveillance, clinical

decision supports, data streaming, genomics and personalized care, consumer-based social media, and support health innovation. However, healthcare and public health sectors generally tend to lag behind in widely adopting and utilizing IT innovations (i.e., BDA) comparing to other sectors (55, 68, 69). For example, full-fledged EMR systems, that collected data directly from physicians from both inpatient and out-patient settings, were experimented and achieved around late 1960's (70). It was only until about ten years ago when EMR started to rise in popularity amongst Canadian physicians (63). Only a handful of stakeholders in Canada have looked deeply into how BDA may apply to them, their stakeholders, or how it fits into their digital health strategy, investment budget, and existing data and analytic infrastructures (66). Even fewer would have the financial leeway and technical supports to embark on a meaningful and sustainable big data implementation.

In 2013, the white paper published by the Canada Health Infoway (CHI) indicated that there has been currently little to no large, enterprise-based, and production-ready BDA solutions existing in Canada (66). In the same year, the chief technology office of the Canadian Medical Association (CMA) described Canadian health care analytics as being in the "teen years" (71, 72). A number of major, deep-rooted, and systemic barriers are summarized below to explain this tendency (26, 55, 66, 69, 73):

Technical barriers

- there is a shortage of data analysts equipped with diversified skillsets, especially those with both technical skills (i.e., data management and programming) and knowledge/experiences pertaining to health systems;

- additional computer science training and computer science personnel may be needed to perform computationally-intensive tasks such as machine learning, programming, data mining, web scraping, and data management;
- in the back end, skilled or additional IT staffs are needed to build, test, and maintain big data systems that require mature networking technology and data entering, securing, and retrieving methods;
- there may be a knowledge gap between decision makers and technical staff as decision makers may lack the in-depth technical understanding of the needs or potential of BDA, while technical staffs may not be aware of the available agendas and rationales at the high level.

Cultural barriers

- there may be a general fear of change by expecting big data implementation to cause sudden and complete overhaul;
- there may be misconceptions regarding big data as unnecessary, impractical, unjustifiably expensive, or inherently unsafe with respect to data privacy and confidentiality;
- certain organization or department branches may not value the evidence-based decision-making process highly;
- some physicians view data digitalization (i.e., EMR) as being impersonal to patients which hinders rapport building;
- digitalization in workplace may require additional training from existing staff members who may not welcome further workloads or responsibilities;

- general public may not be aware of the technicality and needs for advancing/expanding BDA in the health sector, which leads to lackluster supports or initiatives.

Business incentive barriers

- there is a severe lack of evidence quantifying the return on investment (ROI) in both dollar and improvement in patient and population health values of big data implementation in Canadian health sectors;
- there is a lack of known success stories of organizations effectively and sustainably implementing comprehensive BDA platforms for other parties to learn or model from within the Canadian health sectors;
- fiscal incentives of big data implementation are often seen as misaligned with direct benefits to those carrying it out, for example, some physicians view health analytic implementation to be directly beneficial to the health system, payers, and patients, but less so to themselves;
- unsuccessful attempts in big data implementation in the past may be erroneously attributed to the IT itself, rather than the potential issues with implementers, implementation planning and methods, and organization's readiness, resources, and constraints;
- measuring resulting health benefits and ROI is challenging and expensive, and using existing general fiscal metrics for evaluation may be inappropriate by missing the obscure benefits such as operational efficiency and workers' job satisfaction and efficiency;

Infrastructural barriers

- national BDA initiative and standards are difficult to reach since each Canadian jurisdiction is responsible for its own management, organization, and delivery of health care services to its residents;
- within each jurisdiction, health organizations and facilities (i.e., hospitals) are largely independent and self-governing with major decisions usually determined by their own sets of chief executive officers (CEO) and directors;
- health organizations/facilities are complex social environments in which end-users may not comply with the use of recommended IT technologies unless it is convenient and incentivized sufficiently;
- there is a lack of trained professionals with both IT and health care knowledge and skills to help promote and assist the transition;
- in terms of collecting, linking, sharing new or existing big data within and between organizations/departments, the relevant privacy and confidentiality laws may not be clear, available, or appropriate;
- major additional funding and skillsets may be needed to carry out implementation and testing of BDA platforms for a given organization

1.2.4 Future of epidemiology

Mooney *et al.* (2015) predicts that while the core foundation of epidemiology as practiced today is likely to remain the same, this field's potential in maximizing its contribution in understanding and improving health of the population depends on its innovation and integration in big data-related subject matters (25). The CHI depicted that big data knowledge creation can be broken down into three areas including 1) knowledge acquisition, 2) knowledge generation with systemic "automated" analytics, and 3) contextualized clinical and health system analytics

(66) (**Figure 1-2**). Epidemiologists interested in health-related big data and predictive analytics may seek relevant opportunities to contribute in each of these areas in big data knowledge creation. While further training and subspecialization (especially theories and technical skills in computer sciences) may be needed, the core quantitative skillsets learned from traditional epidemiologic training and extensive population health knowledge may serve as an advantage and conceptual foundation for additional quantitative techniques to build on.

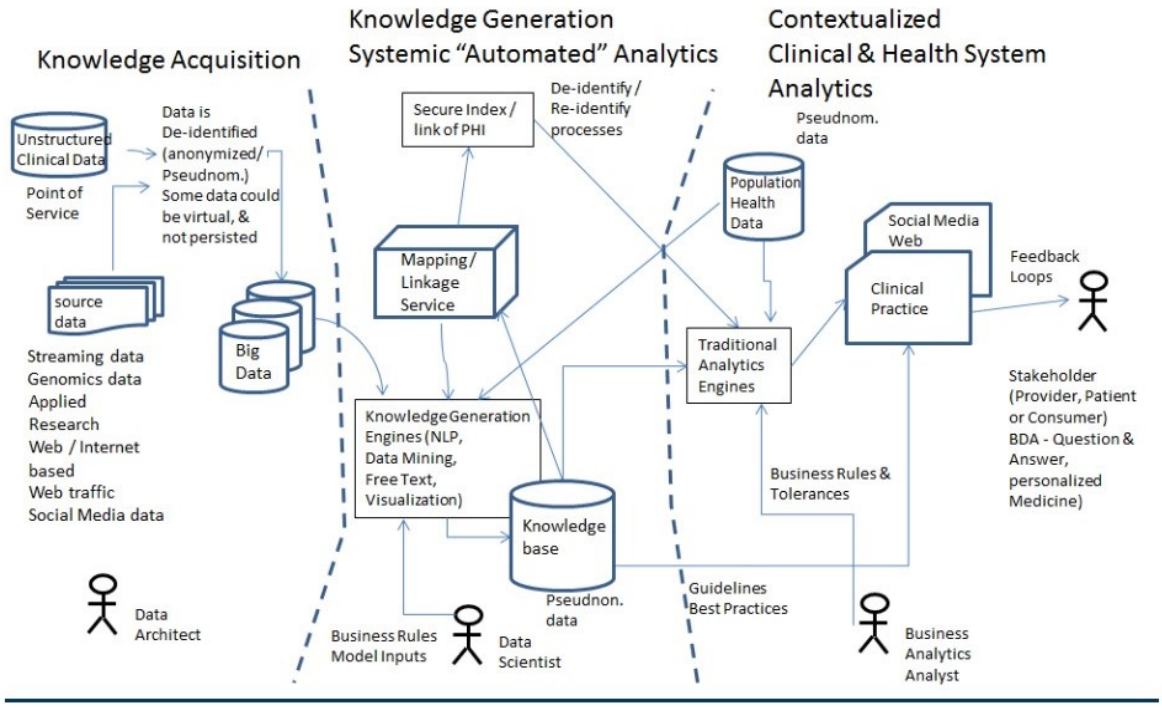


Figure 1-2: Block-diagram showing the components and work flow of a data-driven knowledge-based system (Extracted from (66)).

In 1998, Mervyn Susser, one of the pioneers of epidemiology in the twentieth century, predicted that the traditional epidemiology focusing on single-level risk factor identification will become "less serviceable" compared to the emerging epidemiologic subdisciplines such as genetic epidemiology and global epidemiology that are based on high-volume information systems (74). The big data era has been presenting, and will continue to present, exciting and

novel opportunities for epidemiologists to develop innovative public health research and applications to improve the health of the general population. This global shift of human interactions and dynamics with massive data will likely catalyze a new era of epidemiology that will venture into many uncharted data and analytic territories. Utilizing and linking new types of existing, relevant data may allow epidemiologists to paint a more complete picture of the health of a population and to more effectively monitor, evaluate, and inform health-related decisions and policies in a streamline and timely fashion.

Canada needs to address some of the aforementioned technical, cultural, incentive, and infrastructural barriers in order to fully garner the power and benefits of big data. It is advisable that public health and epidemiologic departments take initiatives to engage in multidisciplinary big data collaborations, rethink workforce development, provide interested epidemiologist trainees adequate training and opportunities, and develop research that explores how public health personnel may deliver unique contribution in big data-related operations (i.e., study design, data validation, new epidemiologic methods for BDA, and incorporation of BDA for health) (75). Momentum is required from all fronts, including individual epidemiologists and data analysts, public health departments, government and organization decision makers, in order to dissolve the existing Canadian barriers in a timely fashion (69).

1.3: General Directions and Research Objectives

This doctoral research aimed to illustrate how big data and ML methods can be applied in public health applications in all three areas of knowledge creation, including knowledge acquisition, knowledge generation, and contextualized clinical/health system analytics. This was achieved by conducting two independent projects, namely *Name- and Location-based Aboriginal Ethnicity Classification* (Study I) and *Sentiment Analysis of Breast Cancer Screening in the*

United States Using Twitter (Study II). There is no direct relationship between the two projects. In Study I, the 1901 Canadian census was collected online and its name and location variables will be used to automatically predict individuals' Aboriginal ethnicity status in Canada. Its significance lies in the general lack of Aboriginal and ethnicity information in Canadian data systems which hinder epidemiologists' ability to investigate health-related information in Aboriginal Canadians. In Study II, tweets mentioning terms related to breast cancer screening were collected via Twitter's REST API. The corresponding sentiment of each tweet is classified using our slightly-modified version of the existing automated sentiment classifier by Hutto and Gilbert (2014) (76). A temporal, geospatial, and thematic analysis is conducted to visualize the sentiment pattern descriptively. The correlation of the aggregated sentiment scores and actual breast cancer screening uptake percentage at the state level in the United States (U.S.) derived from an external national data source is examined.

Chapter 2 – Research Components

2.1 Overview

A considerable portion of big data contains health-related information (77). They capture novel and untapped behavioral, medical, demographic, and public opinion information that can potentially reveal new aspects of population health that are previously-hidden from traditional data sources (78). Traditional epidemiologic methodologies alone may be insufficient to fully utilize health-related big data for predictive analytics, thus tools from other disciplines such as ML should be explored. Traditional epidemiologic methods emphasize making inferences (79). They generally employ classical statistics that are based on probability models of data. In contrast, machine learning emphasizes prediction. While some machine learning algorithms use

probability models of data (similar to classical statistics), other machine learning algorithms do not assume an underlying data model, which are particularly suitable for solving problems involving data that does not fit any a-priori probabilistic assumptions. Currently, epidemiologic exploration of big data and ML research and application in the context of public health is limited. The overall objective of this doctoral thesis is to apply, demonstrate, and examine automated ML approaches on big data to address two important public health issues existing in North America via two independent studies:

- Study I: To predict Aboriginal ethnicity using personal names and residential information in Canada.
- Study II: To conduct a sentiment analysis on breast cancer screening tweets and validate its relationship with breast cancer screening uptakes in the U.S.

2.1.1 Study I: Name- and location-based Aboriginal ethnicity classification

An ethnic group is a socially-defined category referring to people sharing a common cultural heritage, which is comprised of historical memories, cultural practices, and national experiences (80). Ethnicity is linked to personal health which can shape lifestyle behaviors (81, 82) and perceived self-identity (83-85). Like age and sex variables, the variable describing one's ethnicity is considered one of the most fundamental and frequently-studied dimensions in public health (86). Ethnicity (variable) has been used in the construction of a risk factor, confounder, stratifier, and as part of an interaction term in epidemiologic studies. Overall, ethnicity serves the following public health functions: as a 1) risk factor for health outcomes; 2) proxy for unmeasured social factors; 3) marker for unmeasured biological differences; 4) stratifier for vital and health statistics; and 5) catalyst for better delivery of health services (87).

Canada is culturally- and ethnically-diverse, yet abundant evidence demonstrates the insufficiency and inadequacy of ethnicity data in most data sources within the country (88-95). Parsons (2005) described that there is a "profound lack of routine collection of race/ethnicity relevant data in Canada". This compromises the ability to understand and monitor public health issues including health inequality and racial discrimination, potentially leading to culturally-insensitive and -ineffective health promotion, services, and policies. ML methods using personal names to predict ethnicity may be applied to potentially alleviate this ethnicity data gap issue in Canada (96). Naming practices in societies are not random but follow specific cultural and linguistic conventions (88, 93). Through time and space, names preserve a link with particular language used by specific ethnic groups, and these linkages can be established phonologically, morphologically, and semantically (97). To date, no comprehensive study has been conducted to investigate the accuracy of ML-based prediction of individuals' ethnicity using their names in Canada. In particular, Aboriginal Canadians are disproportionately suffering from a wide range of social injustice (98, 99) and health-related problems (100, 101) compared to non-Aboriginal counterparts.

In the national diabetes report, the Public Health Agency of Canada (PHAC) acknowledged the issue of the lack of Aboriginal (and ethnic) identifiers in most administrative databases used for surveillance for diabetes and other chronic diseases (102):

> "*Many data limitations exist for diabetes surveillance for First Nations, Inuit and Métis populations. For example, the inclusion of Aboriginal individuals in national surveys is limited by the geographic coverage of sampling, non-participation, incomplete enumeration of reserves, and exclusion of homeless people. Different survey and sampling methods as well as changes in the criteria for diagnosis of diabetes can also*

interfere with the comparison of survey results between populations and over time. Health administrative data (hospital records, physician billing databases, and provincial/territorial health insurance registries) are often used for diabetes surveillance in the general Canadian population. However, only a limited number of provincial and territorial databases contain Aboriginal identifiers, limiting their use for surveillance for this population."

Currently, there is no study examining if name/location-based ethnicity classification will be applicable and accurate in predicting Aboriginal ethnicity in Canada. This study aims to derive a convenient and timely means to predict Aboriginal classification using name and residential variable data that are commonly collected in most existing Canadian data sources.

2.1.2 Study II: Sentiment analysis on breast cancer screening tweets

In 2016, there were approximately 550 million active Twitter users (those who have created an account and published at least one post) (103). Many Twitter users frequently publish their thoughts, life experiences, opinions, feelings, and other perceived worthwhile information openly in the form of a tweet, a public message of 140 or less characters, in a real-time, streaming fashion (104-106). A considerable portion of tweets is health-related and could potentially be used in public health applications (107, 108). Some of the examples of Twitter-based public health applications include tracking of public awareness of influenza (109), worldwide influenza incidence (110), as well as self-reported mental illnesses (111), medical complaints (112), and safety events by patients (113).

Breast cancer is the most common cancer in women and breast cancer screening is an effective tool for breast cancer control and prevention (114). The United States Preventive

Services Task Force (USPSTF) recommends women aged 50-74 years to attend a mammography every 2 years (115). Evidence of breast cancer screening benefits for women aged 40-49 years is weaker than those in 50-74 years old. Thus for women aged 40-49 years, USPSTF recommends that the frequency of breast cancer screening uptake be decided by the individuals (and their family doctors), but for those who value the potential benefits of cancer screening over potential harms may choose to attend biennial screening (115). However, not all eligible adult women adhere to these breast cancer screening recommendations and their screening uptake behaviours were found to be related to a number of factors such as residential location (116), social class (117), and ethnicity (118, 119). Other determining factors influencing if an eligible woman will seek breast cancer screening tests are related to her perception regarding the level of competence of health professionals, ease of receiving the breast cancer screening procedure, and reasonable wait time to receive the screening procedure and later screening results (120). Women who chose not to attend regular breast cancer screening often cited reasons of being too busy, unaware of breast cancer risks and available screening, anxious/afraid to receive a false or positive result, and worried of the pain/discomfort associated with the procedure (121). Such findings are in accordance with the health belief model (HBM), which states that individual's willingness and action to take a health-related action (i.e., screening) are related to a set of perception-based beliefs including perceived susceptibility, perceived severity, perceived benefits, perceived barriers, cues to action, and self-efficacy (122). Twitter provides a potentially abundant source of perception information from the general public that can be used as a monitoring/surveillance tool for public health.

2.2 Literature Review and Rationale

2.2.1 Study I: Name- and location-based Aboriginal ethnicity classification

Mateos (2007) conducted a systematic review of representative studies, published in or before January 2006, that developed new name-based ethnicity classification methods or evaluated old methods with unknown generalization performance. The 13 included studies aimed to demonstrate a satisfactory rate in classifying individuals into one or a few ethnic groups from the respective general populations (96). These studies shared similar analytical components: source data, reference list, and target population. The datasets for these studies contained both ethnicity and name information, and were generally collected from population administrative files, health registries, or surveys (96). A reference list is an exhaustive list of personal names, each with a true ethnic classification. The features used to train the ML algorithms among these studies were primarily name components (such as first, middle, and last name). For those studies using the reference list to train the ML algorithms, individuals from the target population were then classified into ethnic groupings (96). Generalization performance measures, such as accuracy, sensitivity, and specificity, were used to evaluate the degree of agreement between the predicted and true ethnicity classification.

Among the 13 included studies, sensitivity ranged between 0.67 and 0.95, specificity between 0.8 and 1.0, PPV between 0.7 and 0.96, and NPV between 0.96 and 1.0 (96). The authors in these studies generally interpreted their results as satisfactory, serving as a validation that name-based ethnicity classification is adequate to ascribe the ethnicity variable for a number of ethnic groups (96). Only three of the 13 included studies in Mateos' review (2007) were conducted in Canada, namely Coldman *et al.* (1988) (123), Choi *et al.* (1993) (124), and Sheth *et*

al. (1999) (125). An additional, similar study was published more recently by Shah *et al.* (2010) (126). They examined either one (Chinese only) (123, 124) or a few ethnic categories (South Asian and Chinese only) (125, 126). Three of them (124-126) used only surname while one used forename, surname, and middle name (123). Only one (125) used national data while the other three used individual provincial data with limited sample size ($n < 5,000$) (123, 124) or larger sample size ($n>1,000,000$) (126).

Overall, there is a scarcity of research and application exploring name/location-based classification methods to predict one's ethnicity. In addition, despite the health inequities experienced by the Aboriginal Canadians and recent studies showing residential information improved name-based ethnicity prediction (127, 128), no study has examined name and location features to predict Aboriginal ethnicity in Canada. Thus, the main objective of this study is to evaluate the generalization performance of using name only and name plus location information to derive a comprehensive set of features in order to predict Aboriginal ethnicity status in Canada using a large dataset.

2.2.2 Study II: Sentiment analysis on breast cancer screening tweets

Internet has become increasingly interactive via the plethora of online social platforms of blogs, forums, commenting sections, chats, and social media. Patients and the general public no longer merely consume web information passively. They now actively engage in discussions and generate a massive amount of timely, accessible, and unfiltered unstructured data. Online patient perception (or opinions) can be a major source of collective intelligence for evaluating the performance of hospitals such as the timeliness and quality of care, and effectiveness of public health interventions and policies (129, 130). Twitter has hundreds of millions of active users, regularly broadcasting their thoughts and opinions on diverse topics (131). Various aggregate

analyses of opinions using Twitter have shown success in modeling phenomena such as elections (132), presidential job approval rating (133), stock market (134), and consumer confidence (135).

Much of the previous research efforts on health-related applications using Twitter data have focused on disease/syndromic (109-111, 136) and natural disaster surveillance, (137, 138) hence new terms "infodemiology" and "infoveillance" were coined (139). Much less research has focused its attention on exploring the applications of Twitter users' perception towards clinical interventions. Salathé and Khandelwal (2011) conducted a sentiment analysis measuring the temporal and geographic patterns of sentiment towards a novel Hemagglutinin Type 1 and Neuraminidase Type 1 (H1N1, such as influenza A) vaccine in the U.S. They found a strong correlation between sentiments expressed online and Centre of Disease Control and Prevention (CDC)-estimated vaccination rates by region (140). They also found that communities generally tended to be dominated by either positive or negative sentiments towards the vaccine. Another Twitter opinion-based health research was conducted by Myslin *et al.* (2013), examining the sentiment of tobacco products with a focus on new and emerging products such as hookah and electronic cigarettes. Novel insights were derived from the observed high prevalence of positive sentiment towards tobacco products. The authors stated the positive sentiment was correlated in complex ways with social image, personal experience, and recently popular products such as hookah and electronic cigarettes which were manually annotated in content analysis using a different set of smoking-related tweets (141).

Twitter has become a popular channel to connect people interested in cancer-related topics. Professional medical societies such as the American Society of Clinical Oncology (ASCO) uses Twitter to report clinical news and publications, to discuss treatment issues, and to

facilitate a broader dialogue amongst physicians and health professionals (142). Individual physicians can directly reach their patients to update therapeutic advancements, to answer disease-related questions, or to provide medical advice or reminders (143). Cancer patients can help support and educate each other, and share their experiences and opinions about their conditions and treatments received (144, 145). The general public can also use Twitter to show support for public awareness and fund-raising events, such as breast cancer awareness month (146). It should be clear from these diverse examples that Twitter is a rich source of data for understanding the value, perspective, and interaction between people with different cancer-related interests and needs. Particularly, Twitter can be a great source of information to understand people's sentiments about breast cancer screening tests, to identify social or infrastructural barriers that make using the screening difficult, and to inform decision-makers when, where, and to whom resources should be allocated.

Research on Twitter's sentiment on breast cancer screening is currently lacking. To date, the only study addressing this topic examined a very small sample of mammogram-related tweets (n = 271) using just a simple search term ("mammogram") for breast cancer screening (147). Thus, the objective of this study is to conduct a comprehensive analysis on breast cancer screening tweets to visualize the patterns in time, space, distribution, and theme of the sentiment expressed towards breast cancer screening among U.S. citizens. In addition, this study also examined the correlation between breast cancer screening sentiment (from Twitter) and actual breast cancer screening uptake (from an external national database) in the U.S.

2.3 Specific Objectives

2.3.1 Study I: Name- and location-based Aboriginal ethnicity classification

- To conduct a literature review to describe the current ethnicity data gap issue common among health-related data systems within Canada and corresponding recommendations (as position paper in Chapter 3);
- To use the 1901 Canadian census to train and test name- and location-based ethnicity classifiers for the following Aboriginal categories: 1) Aboriginal (all-inclusive), 2) First Nations, Métis, and Inuit, 3) major Aboriginal language groups, and 4) major Aboriginal tribal groups. The accuracy of the classifiers will be evaluated against a-priori generalization performance metrics (as publication manuscript in Chapter 4).

2.3.2 Study II: Sentiment analysis on breast cancer screening tweets

- To use breast cancer screening-related tweets collected between September 2014 and April 2015 to visualize the temporal, geospatial, and thematic patterns of sentiment (positive, negative, and neutral) towards breast cancer screening tests in the U.S. (as published article in Chapter 5);
- To examine the relationship between sentiment of breast cancer screening and breast cancer screening uptake by states using an external national database (as published article in Chapter 5).

2.4 Research Methods

2.4.1 Study I: Name- and location-based Aboriginal ethnicity classification

The following described research methods correspond to the publication manuscript to be presented in Chapter 4.

2.4.1a Research questions

- Question 1 (S1-Q1): What is the generalization performance (in accuracy, ROC, sensitivity, specificity, PPV, and NPV) of the ML classifiers for predicting Aboriginal (all-inclusive) status based on name alone and on both name and residential location?
- Question 2 (S1-Q2): What is the generalization performance of the ML classifiers for predicting subgroups of Aboriginal (First Nations, Métis, and Inuit; major Aboriginal language groups, and major Aboriginal tribal groups) status based on name alone and on both name and residential location?
- Supplementary question 1 (S1-SQ1): What is the generalization performance of the ML classifiers for predicting non-Aboriginal ethnicities (Chinese, English, French, Irish, Italian, Japanese, Russian, Scottish) based on name alone and on both name and residential location?

2.4.1b Data source and data processing

The 1901 Canadian census, which officially began on March 31st, serves as the sole data source for this study. The 1901 census employed a total of 351 commissioners to coordinate data collection throughout the country (148). There were 8,800 enumerators visited 206 census districts, divided into 3,204 sub-districts consisted of cities, towns, groups of townships, Indian reserves, and other less well-defined areas. Enumerators collected census information from

5,371,315 individuals distributed as British Columbia (178,657), Manitoba (255,211), New Brunswick (331,120), Nova Scotia (459,574), Ontario (2,182,947), Prince Edward Island (103,259), Quebec (1,648,898), and Territories (211,649) (148). Census can be considered as big data with its large sampling coverage, value of representing the entire national population, and since name and ethnicity fields in the 1901 census were entered in free-text format, it represents a low level of veracity in the data. Analyzing this census data is considered a big data problem since it contains large volume of cases and unstructured textual inputs for name and ethnicity variables, thus signifying the *volume* and *veracity* aspects of big data.

The 1901 census was made publicly accessible from multiple websites (149-151). User agreements were followed and consent for using the 1901 census content for this research was obtained. A web indexing codebase was developed in Python 2.7, along with corresponding libraries (such as urllib2, re, and beautifulsoup). The 1901 census provides all the required individual-level data fields including name, residential location, and ethnicity (152). It is an appropriate dataset for this research since the frequency of interracial marriages and marriage-related name changes were less common in the past century thus preserving the association between name and ethnicity (153). At the same time, this census is only three to four generations old, thus its overall generalizability to modern time should still hold true.

Individuals missing either name and/or ethnicity information were removed from the dataset. Since both the name and ethnicity data fields are in unstructured format, data cleaning and processing were performed. For the name field, abbreviated titles, such as Dr, Esq, Hon, Jr, Mr, Mrs, Ms, Messrs, Mmes, Msgr, Prof, Rev, Rt Hon, Sr, and St, were stripped away. Likewise, any numbers or nonsensical punctuations, such as “?”, “:”, “*”, “%”, “^”, “&”, “#”, “@”, “$”, and “!”, were removed from the name field. The name field was then split into two name fields:

first and last name. All the letters belonging to the last connected sequence of alphabets were extracted as the “last name” field, while the rest of the text were extracted as the “first name” field. All of the above steps were carried out using Python 2.7 and its regular expression library.

The ethnicity field initially contained 6,763 unique entries. The reason for this large volume of entries was due to the unstructured, free-form text entering format, which allowed for abbreviated, alternatively-spelled, and misspelled forms of the same ethnicity. The standardization of similar ethnicities from different forms was examined manually, and then a set of heuristic rules were used to categorize the designated ethnicity label. Using regular expression, any entries containing nonsensical punctuations such as “?” and “%” and/or with no alphabetical letters were replaced with an empty entry, and the record was removed from the dataset. Misspelling and alternative spelling (i.e., French spelling) forms for a particular ethnicity were identified and recategorized. For example, “Métisse” and “Métise” were recategorized into “Métis”. Individuals with “First Nations”, “Métis”, and “Inuit” were also given an “Aboriginal” (all-inclusive) ethnicity status. First Nations individuals were further broken down into finer groups: major Aboriginal language (Algonquian, Athapaskan, Iroquoian, Kootenay, Salish, Siouan, Tsimshian, and Wakashan) and major Aboriginal tribal (Algonquin, Blackfoot, Cree, Iroquois, Micmac, Mohawk, and Ojibwa) groups. Each of these subgrouping contains further individual ethnicity entries. For example, the Algonquian language group consists of ethnicity entries such as “Algonquin”, “Blackfoot”, “Cree”, “Ojibway”, “Malecite”, and others. The heuristic rule sets were constructed using an iterative cross-checking method with various referencing resources (154-158). The frequency of each ethnic category will be presented in Chapter 4.

2.4.1c Feature generation

The aforementioned first and last names gave rise to 15 name-related features. These name features were based on single characters, substrings of characters, entire name entities, and phonetic representation. The construction of these features will be discussed at length in Chapter 4. The residential location feature was derived from the sub-district where the census information was taken from individuals, The default entry contains four location segments at various granularity, for example, “B, Glengarry, Ontario, Canada”. To avoid low number of small regions, the final residential location feature striped away the finest regional description, for example, “B, Glengarry, Ontario, Canada” into “Glengarry, Ontario, Canada”.

2.4.1d Analytic approach

The ML process was done on the full 1901 census (N=5,031,794) or a 10% random subsample (N=500,000) in order to utilize all available data and lower computation time, respectively. Data was randomly split 50:50 into training and validation sets. One-versus-the-rest binary classification was conducted for each ethnic category (i.e., “Aboriginal” versus “non-Aboriginal”; “First Nations” versus “non-First Nations”, and “Cree” versus “non-Cree”). The process of machine learning took place within the training set where three ML algorithms, including (regularized) LR, SVM, and DT, were used to associate feature values and corresponding ethnicity labels. More detailed description of the ML algorithms will be available in Chapter 4. Either the full set of name and residential location features or name features only was examined for each training. Once the ML model was trained, it was used to predict ethnicity of the sample within the validation set. Its generalization performance, or degree of agreement/disagreement between the predicted and the true ethnicity values, was carried out using a widely accepted set of measurement metrics including accuracy, ROC, sensitivity,

specificity, PPV, and NPV. Results will be presented in table format (Chapter 4) showing the generalization performance metrics by different ethnic groupings of Aboriginal status, ML classifiers, sample sizes, and feature sets. In addition, the most informative features will be generated using the DT classifiers.

To evaluate the applicability of the resulting ML models, simulated disease/population scenarios will be set up in an demonstrative application (Chapter 4). Hypothetical population size, proportion of ethnic composition, and proportion of diseased individuals were preset to cover a wide range of possibilities. From these, the underlying "true" disease prevalence and corresponding 95% confidence intervals (CI) can be calculated. From the aforementioned generalization performance results, predicted disease cases and hence predicted disease prevalence can be further obtained. Comparisons between the predicted disease prevalence and true disease prevalence can be made to estimate the degree of error may be made by the obtained ML classifiers in a more relevant public health setting. All of the analyses of this study was performed using Microsoft Excel and Python 2.7 with its numerical computation, ML, and data management libraries such as numpy (159), scikit-learn (160), and pandas (161), correspondingly.

2.4.2 Study II: Sentiment analysis on breast cancer screening tweets

The following described research methods correspond to the published article (162) to be presented in Chapter 5.

2.4.2a Research questions

- Question 1 (S2-Q1): What are the temporal patterns in breast cancer screening tweet frequency and sentiment in the U.S. over the study period?

- Question 2 (S2-Q2): What are the geospatial patterns in breast cancer screening tweet frequency and sentiment in the U.S. over the study period?
- Question 3 (S2-Q3): What are the thematic patterns in breast cancer screening tweet frequency and sentiment in the U.S. over the study period?
- Question 4 (S2-Q4): Are there any significant ($p<0.05$) associations between breast cancer screening sentiment scores and uptake of breast cancer screening in the U.S. at the state level?
- Supplementary question 1 (S2-SQ1): What are the temporal, geospatial, and thematic patterns in breast cancer screening tweet frequency and sentiment in Canada over the study period?

2.4.2b Data source and data processing

Twitter serves as an unconventional big data source for epidemiologic research due to its large volume, real-time production, and open public access. In 2014, Twitter had 255 million monthly active users worldwide (163). There were approximately 500 million tweets sent per day, and 46% of active Twitter users tweeted at least once a day. One percent of random tweets were made available for public access. For this study, breast cancer screening-related tweets were collected for about eight months between 17th September 2014 and 10th May 2015, via the Twitter REST API using Python 2.7 and its Tweepy library (164). Tweets mentioning at least one of the following breast cancer screening-related terms were included: "mammogram", "mammography", "breast imaging", "breast screening", "breast mri", "breast ultrasound", "breast self-exam", "breast examination", "breast exam", "breast lump", and their corresponding hashtag forms such as "#breastscreening" and "#breastexam". The extracted information from each included tweet included user name, user profile description, time of tweet, tweet itself, retweet

status, user mentions, hashtags, and geographical information including user-described location field and user-enabled latitude and longitude coordinates (or GPS) field (165).

After examining the geographic distribution of the collected Twitter data, a post hoc decision was made to focus the study only on the U.S. (rather than North America as a whole) since the amount of breast cancer screening tweets from Canada was insufficient for meaningful analyses. Hence, non-U.S. tweets were excluded for further analysis. The textual content of each included tweet was processed by stripping away the retweet tag ("RT"), hashtag symbol ("#"), user-mentioned tag ("@"), and Uniform Resource Location (URL) links. For the fields that describe user's location while the tweet was publicized, there were two possible location fields. The user-described location field allows Twitter users to write a description of their current location, and the GPS field automatically recorded the latitude and longitude coordinates from the electronic device. If both location inputs were available, the more precise GPS location was used, otherwise the user-described location was used. If existed, user-described location was converted into GPS coordinates using Python module Geocoder (by accessing MapQuest) (166). The location information was then standardized by reverse-geocoding the coordinates into the corresponding country, state, county, and city. This Twitter-provided information will be used to explore the temporal, geospatial, and thematic patterns of sentiment towards breast cancer screening in the U.S.

For the association analysis between breast cancer screening tweets and breast cancer screening uptake, an external data source was used to provide the information on uptake behaviours in the U.S. population. The Behavioral Risk Factor Surveillance System (BRFSS) is a telephone survey conducted by the CDC annually since 1984 to collect health-related risk behaviours, chronic health conditions, and use of preventive services such as breast cancer

screenings (167). The anonymized BRFSS 2014 (for calendar year 2014) was made publicly downloadable by the CDC (168), which was used for this study. Interested individual-level fields were extracted from the BRFSS 2014 including 1) most recent mammogram received, 2) most recent clinical breast exam (CBE) received, 3) highest education achieved, 4) self-reported general health status, and 5) self-reported race. The collection of education, general health, and race variables was done in order to adjust for potential confounding in the analysis.

A number of data processing took place in the BRFSS data. Since the scientific evidence for the benefits from attending regular breast cancer screening and professional recommendations for attending such practice were strongest for women aged 40 to 74 (115), women aged younger than 40 or older than 74 and men were removed from the data. Each of the five aforementioned variable fields contained multiple possible response categories. In order to eliminate possible low response frequency and increase interpretability, dichotomization of each variable was carried out. The recoded variables included 1) mammogram received within the last two years (1 – yes and 0 – no), 2) CBE received within the last two years (1 – yes and 0 – no), 3) highest education achieved (1 – have at least some college education, 0 – do not have any college education), 4) general health (1 – good, very good, or excellent, 0 – fair or poor), and 5) race (1 – non-Hispanic white only, 0 – all others).

2.4.2c VADER sentiment classifier and its modifications

Each of the collected tweet needs to be classified to a corresponding sentiment (i.e., positive, negative, and neutral). Manually labelling each tweet by human raters will be a labor-intensive and inefficient, thus a ML method is used. There are currently a number of existing automated sentiment classifiers (169), including the Linguistic Inquiry and Word Count (LIWC), General Inquirer (GI), Affective Norms for English Words (ANEW), SentiWordNet (SWN),

SenticNet (SCN), Word-Sense Disambiguation (WSD), and Hu-Liu-2004. However, their generalization performance in classifying sentiment in tweets may not be optimal since they were not developed specifically for microblogging platforms populated with short messages, such as tweets). Hutto and Gilbert (2014) have developed and made publically available their sentiment classifier, called Valence Aware Dictionary for sEntiment Reasoning (VADER), targeting microblogging platforms (169). They have employed various novel sentiment signals' detection methods such as the inclusion of acronyms, slangs, emoticons as part of their lexical dictionary. They then empirically validated and rated each lexical feature over a continuum of sentiment range by multiple tested and trained raters. Afterward, they combined these lexical features with consideration for five general rules that embody grammatical and syntactical conventions for expressing and emphasizing sentiment intensity (169). For example, an exclamation point "!" (relative to a period or no punctuation at all) increased the sentiment intensity by a formulated magnitude.

The VADER classifier examines a text (such as a tweet) and produces four sentiment scores: 1) neutral, 2) positive, 3) negative, and 4) composite scores. The neutral, positive, and negative scores correspond to the proportion of text containing a particular sentiment polarity. For example, a positive score of 0.5 indicates that 50% of the words in a tweet contain positive sentiment while the other 50% contain either neutral or negative sentiment. The composite score is computed by summing the sentiment intensity score of each word from the text that has a match with the VADER lexicon, adjusted with grammatical and syntactical rules, and then normalized to be between -1.0 (most negative) and +1.0 (most positive). The composite score was categorized into neutral (-0.3 to +0.3), positive (>+0.3 to +1.0), and negative (-1.0 to <-0.3). The composite score can be used as a single unidimensional measure of sentiment. Hutto and

Gilbert (2014) concluded the VADER classifier considerably outperformed the seven aforementioned sentiment classifiers. The VADER classifier achieved a 0.99 precision, 0.94 recall, and 0.96 F1 score, which were comparable to human accuracy (169).

A validation was conducted to assess the generalization performance of the VADER classifier on the collected tweets specific about breast cancer screening. A random subset of 250 tweets was drawn and each was given a subjective sentiment label (neutral, positive, or negative) by KW (as independent gold standard). These tweets were then processed by the VADER classifier, and each composite score was converted into a neutral, positive, or negative label. The accuracy was suboptimal at <40%. The overarching reason was identified. In the original VADER lexical dictionary, the lexicon "cancer" contained a highly negative sentiment value (-3.4). This resulted in VADER classifier overly assigned highly negative composite sentiment score to virtually all tweets since they were related to breast cancer by default. Similarly, other words including "die", "died", and "death" containing highly negative default sentiment values (i.e., -2.9, -2.6, and -2.9, respectively) were identified, yet these lexicons often appeared in the collected tweets without any underlying positive or negative connotation. The effect on sentiment classification accuracy was examined by removing these lexicons from the original lexical dictionary, resulting in more favorable accuracy (77.2%). This modified version of the VADER sentiment classifier was used throughout the analysis of this study.

2.4.2d Analytic approach

There was a total of 54,664 breast cancer screening-related tweets collected from the U.S. with state-level geographic information, over the study period. The analyses can be divided into 1) descriptive analysis (exploring temporal, geospatial, and thematic patterns) and 2) hypothesis-driven analysis (testing the association between screening sentiment and screening uptake). The

analysis of this Twitter data set is considered a big data problem since this data contains real-time and unstructured textual data that signifies the *velocity* and *veracity* aspects of big data. Also, while our study only examined 8 months of tweets pertaining to breast cancer screening, Twitter houses a much larger amount of tweets in the back end with all discussed topics continuously and simultaneously, thus signifying the massive *volume* possible to be utilized.

For temporal patterns, the tweet volume and national average composite (VADER) sentiment score will be plotted against time. For geospatial patterns, basic point maps, hot spot maps, and quintile maps will be generated. Basic point maps will not only show how (breast cancer screening-related) tweets were distributed, but also how neutral, positive, and negative tweets were distributed across the U.S. Hot spot maps will identify statistically-significantly high or low regions with respect to their sentiment towards breast cancer screening. This will be conducted by comparing local neighbouring region's sentiment values to the global sentiment values using the Getis-Ord Gi* statistics (170), described in more details in Chapter 5. Three quintile maps will be generated. The first quintile map will be created by ranking the average composite sentiment score by each state. The second quintile map will be based on percent women aged 40 years and above receiving a mammogram within the last two years by state, and the third quintile map will be based on percent women aged 40 years and above receiving a CBE within the last years by state. The general trends (such as similarity in high and low quintile regions) between these quintile maps will be examined qualitatively. For thematic patterns, word-cloud analysis was conducted. Word-clouds are big data visualization that presents highest frequency words (while omitting non-informative common words such as "the", "it", and "what"). Word-clouds can be generated using all tweets, as well as only positive or negative tweets. They can also be made using tweets from all or specific regions only. This allows a

possible subjective interpretation of the main themes frequently discussed amongst individuals who favor or disfavor breast cancer screening in specific region.

The possible association between breast cancer screening sentiment and beast cancer screening uptake will be examined at the state level, since these two pieces of information came from two independent data sources and data linking at the individual level was not possible. Each of the four (neutral, positive, negative, and composite) VADER scores will be averaged at the state level. Likewise, the percentages of women aged 40 years and above receiving a mammogram and receiving a CBE within the last two years will be computed at the state level. The state-level recent mammogram and recent CBE percentages will be the outcome variables, while the state-level averages of VADER sentiment scores will be the predictor variables. Since women's education, general health status, and race may influence their likelihood in seeking breast cancer screening (117, 118, 120), these variables were included to account for possible confounding. Since the possible outcome variables' value ranges between 0% to 100%, multivariate beta regression was used to test for significant association at $p<0.05$ (more details in Chapter 5).

Data cleaning and processing, as well as the temporal and thematic descriptive analyses were carried out and visualized using Python 2.7 and libraries (including re (or regular expression), scikit-learn, numpy, matplotlib, wordcloud, and pandas); automated sentiment classification was based on a modified version of Python's library vadersentiment 0.5; map generation and hot spot analysis were done by Arcgis 10.1; and beta regression modeling was carried out using Stata 14.

Chapter 3 – Position Paper: Concerning Data Inadequacy on Ethnicity and Race in Canada - Description and Recommendations

3.1 Defining Ethnicity and Race

The concepts of ethnicity and race are related but not identical, thus they should not be used interchangeably. These concepts play an important role in shaping individuals' self-identity and lifestyle behaviours, making them important information for public health. This position paper will 1) describe the working definitions of ethnicity and race, 2) illustrate the importance of ethnicity and race in public health and epidemiology, 3) discuss the issue of inadequate ethnicity and race data in Canada, and 4) provide corresponding recommendations.

3.1.1 Definition of race

Race is generally defined by a set of physical hereditary traits such as skin colour, facial features, and hair texture which represent ancestral adaptations to different environments. The physical traits that distinguish one race from another are attributable to small genetic differences (171). For example, there is only a 0.005% variation in human genome between a Black African and a White Nordic (172). Hence, some scientists believe that the concept of "human race" has little biological basis (171, 173). The rest of this position paper will refer to the term race as perceived race, rather than a construct based on biologic inheritance. Race has been widely used as a taxonomic categorization of people in historical, economic, and political processes. Winant (2000) described that race conceptualization "began to take shape with the rise of a world political economy. The onset of global economic integration, the dawn of seaborne empire, the conquest of the Americas, and the rise of the Atlantic slave trade, were all key elements in the genealogy of race" (174). To this day, race continues to be a common means to categorize

populations into subgroups, despite the controversy in its biological underpinning, in most countries, including Canada.

3.1.2 Definition of ethnicity

Unlike race, ethnicity is not defined by mere physical traits. Ethnic identity is formed through an internal, transient, and multifaceted process of self-identification in relation to ancestry, kinship, religion, language, world views, cultural traditions, shared territory and nationality, and political and religious beliefs that are deeply rooted and differentiated by different ethnic groups (97, 172). An ethnic group maintains those characteristics from generation to generation with a common history and origin and a sense of self-identification with the group, where members of the group generally share common ancestry, distinctive features in their way of life, and national, political, and social experiences (175). Thus, ethnicity is closely tied to one's perceived self-identity (84, 85) and lifestyle behaviours (81-83). The sense of ethnic belonging is developed over time through a subjective belief of shared genealogy, without the necessary existence of biological or physical similarities (97). The belief of shared genealogy can be actual or presumed. For example, if an individual believes himself or herself to be descended from certain ethnic group and wants to be associated with that group, then he or she is considered a member of that ethnic group (176).

3.2 Why are Ethnicity and Race Data Important in Epidemiology?

3.2.1 Ethnicity and race in epidemiologic research

Our ability to classify entities, including living beings and inanimate objects, into subcategories with similar features serves our essential needs of making sense of this physical and social world. Social classification of human groups was among the most influential human

classification systems (177). Through time, ethnicity and race became the key dimensions of social classification of humans. Information on ethnicity and race of a population generally came from surveys and demographic, administrative, health risk, and health status databases (178). Using available data, epidemiologists have found ethnicity and race to be significant indicators for many health-related dimensions, including self-rated health, chronic conditions, well-being, health expectancy, access to health care providers and services, and environmental health (179). The complex, multidimensional, and intertwined pathways between ethnicity/race and health outcomes are summarized in **Figure 3-1**, by Bay Area Regional Health Inequities Initiative (BARHII) (180). Ethnicity and race can interact and influence downstream behavioural, physical, environmental, cultural, socio-economic, political, historical, and legal factors that could subsequently impact health outcomes. Comstock *et al.* (2004) conducted a comprehensive literature review on studies published in the American Journal of Epidemiology and American Journal of Public Health between 1996 and 1999. From the included studies (N=1,198), the majority of them (N=919 or 76.7%) used ethnicity and/or race as a variable in the analysis. The ethnicity and race variables were used to serve a multitude of analytic roles, including risk factor, confounder, stratifier, covariate, adjusted factor, and part of an interaction term (86). Over half of the included studies (55.3%) in Comstock *et al.* (2004) review have found statistically significant findings related to ethnicity or race. Overall, the ethnicity and race variables enable epidemiologists to achieve various overarching public health goals including 1) suggesting leads about etiology, 2) understanding the roles of, and interactions between, social, genetic and environmental factors, 3) assessing how the conceptualization of risk factors, symptoms, and diseases may differ by ethnicity or race, so that public health interventions may be better tailored to specific groups, 4) considering whether biology (i.e., disease mechanisms and drug

metabolism) may be different within and between ethnic and racial groups, and 5) identifying subgroups that may receive unequal prevention, screening, or treatment, so that public health policies and interventions may be better tailored and targeted (181).

3.2.2 Ethnicity (including Aboriginal status) and race as key stratifiers for health inequity

Health inequity has become a prominent health concept indicating the unequal and unfair distribution of disease burden and health resources between subgroups of a population. As large and persistent health inequities may suggest ineffective components within the health and social justice systems. One of the most formidable public health challenges Canada currently faces is to understand and "take action to improve equitable access to the conditions that affect the health of all people living in Canada" (182). Over the past decade, a momentum has been building in Canada and internationally towards identifying and addressing inequities in health care (such as access, quality, and outcomes) and health status (such as well-being and disease incidence and prevalence) (183). In 2003, the World Health Organization (WHO) Regional Office for Europe identified key socioeconomic factors responsible for the growing health inequities between and within countries, known as the social determinants of health (SDoH) (184). To better fit its context, Canada further refined and developed its own set of SDoH which includes ethnicity and race (together will be labeled as "ethnicity/race" for the rest of this paper), as well as Aboriginal status (185). Aboriginal status is essentially a subcategory of ethnicity, however the deep-rooted, systemic, and pervasive health and socioeconomic disadvantages disproportionately experienced by Aboriginals justified an entirely separate category.

On March 22nd 2016, the Canadian Institute for Health Information (CIHI) has facilitated a pan-Canadian dialogue to advance the measurement of equity in health care. The panel with diverse stakeholders has gone through a series of consensus-building exercises and identified the

following core stratifiers as highest priority for measuring equity in health care in Canada: ethnicity/race, Aboriginal status, as well as age, sex, geographic location, income, education (183). The key rationale in selecting these stratifiers was due to the large group-level differences in health status and health care affecting the majority of Canadians. For example, Canadians of visible minority (such as non-Caucasian, non-Aboriginal racial groups including Black, South Asian, Chinese, and others (118)) or Aboriginality (such as First Nations, Métis, and Inuit) tend to have higher prevalence and worse outcomes for a large number of health conditions (186-188), yet they are also less likely to utilize or have adequate access to available health services compared to the rest of Canadians (189-191). Evidently from this high priority identified by the pan-Canadian stakeholder group, the collection and availability of accurate, reliable, timely, and sufficiently-large data on ethnicity (including Aboriginal status) and race are indispensable for meaningful epidemiologic research and effective health policy and program development in Canada.

3.2.3 Benefits of collecting good quality ethnicity and race data

While the framework in **Figure 3-1** depicts the possible relational pathways between race and health outcomes, it can only serve as a general roadmap. In reality, some pathways may be more important and relevant than others for a particular geographic region. Chinese women tended to have lower breast cancer screening levels across Canada (118). The specific reasons for their low screening uptake may be different from region to region. For example, Chinese women in one particular city may not be aware of the necessity to screen, while Chinese women from another city may be aware but access barriers (such as lack of translators or far distance from cancer screening centers) deter them from such practice. Both cities could be from the same province which may have insufficient funding and existing infrastructure to promote and provide

adequate screening-related services (such as Chinese-English translators) targeting Chinese women. Thus, having accurate, reliable, and continuous ethnicity/race data at various regional and organizational levels is important to show ethnically/racially-specific health patterns in the population of interest in a high-resolution and timely fashion. Specific geographic divisions (such as community, town, city, and province) with their own sets of population composition, infrastructure, agendas, needs, available resources, and constraints should identify their own specific factors, pathways, and interrelationships at play. When good quality and high-resolution ethnicity/race information is available, incidence and prevalence statistics can be stratified by ethnic/racial groups at a desired level. Large differences in disease occurrence and health burden between ethnic groups suggest potential inequities in quality, access, and delivery of health services that need to be further investigated at the system and local levels (192, 193). Available ethnicity and race data in various organizations (such as hospitals and clinics) will enable such in-depth investigation by stratifying utilization and service quality data by ethnic/racial groups. Decision makers (such as government policy-makers and community health planners) will then be more in tune with the characteristics and needs of the ethnic/racial groups in their communities, and culturally-appropriate policies and interventions are more likely to be developed, evaluated, and maintained (87, 194).

3.3 Issues and Impacts Pertaining to Inadequate Ethnicity and Race Data in Canada

With over 200 ethnic origins, Canada is one of the most multicultural countries in the world (195, 196). The most frequently self-reported ethnic origins in Census 2006, either alone or with other origins, were Canadian, English, French, Scottish, Irish, German, Italian, Chinese, North American Indian and Ukrainian. In 2006, 16.2% and 4.0% of the Canadian population

were visible minority and Aboriginal, respectively (195, 197). Between 2001 and 2006, Canada's visible minority population grew by 27.2%, or more than five times the growth rate of the rest of Canadians (195). Between 1996 and 2006, Aboriginal population grew by 45.0%, or nearly six times compared to the rest of Canadians (197).

3.3.1 Inadequate ethnicity/race information in survey data

Despite the multi-ethnic/multi-racial nature of Canada and its recognition of the importance of ethnicity/race information in research and organization settings, data sources providing ethnicity and race information about the Canadian population are scarce and limited. Many national or regional surveys containing ethnicity (including Aboriginal status) or race questions have been discontinued, including the National Population Health Survey (NPHS) (1994-2011) (198), Ethnic Diversity Survey (EDS) (one time in 2002) (199), National Longitudinal Survey of Children and Youth (NLSCY) (1994-2009) (200), and Aboriginal Children's Survey (ACS) (one time in 2006) (201). Four major active national surveys containing ethnicity/race questions are listed in **Table 3-1**, including the Canadian Census, Canadian Community Health Survey (CCHS), Canadian Health Measures Survey (CHMS), and Aboriginal Peoples Survey (APS). Each of these surveys suffered from various limitations such as infrequent survey cycles, lack of standards and comparable definitions and categorization of ethnicity/race, insufficient sample size (especially for small communities), ethnicity/race misclassification, limited population coverage, and low response rates (**Table 3-1**). In terms of coverage, the CCHS, CHMS, and APS excluded Aboriginals living on reserves or in Aboriginal settlements, as well as children below certain age cutoffs. While the coverage of Census is large and representative to the entire Canadian population, its pitfalls pertaining to infrequent survey cycles, undercounting certain population subgroups, lack of consistency between recent cycles

(such as the abandonment of mandatory long-form in Census 2011), and general lack of extensive health-related questions vastly limit its utility in epidemiologic research. While the CCHS and CHMS contained extensive health-related questions, they suffered from poor response rates which called for caution for potential selection bias.

The Centre for Education Statistics (CES) has scanned the availability and data quality of Aboriginal status data across Canada (202). They have identified the following limitations: 1) except the Census and 1991/2001 APS, no survey was done on Aboriginal reserves; 2) only the Census, APS and ACS were considered as major sources of data on Aboriginals, yet the ACS was conducted only at a one-time basis in 2008 and APS's sampling frame was selected from NHS respondents' unvalidated self-reported Aboriginal status; 3) many of these surveys lacked sufficient sample size to provide reliable estimates for the overall Aboriginal population. Thus, high-resolution analyses (such as by children, specific communities, Aboriginal subgroups) were virtually impossible, 4) most of these surveys were based on Aboriginal self-identification without validation; 5) none of the survey was representative to the entire Aboriginal population of Canada; 6) lack of consistency regarding the measured characteristics between surveys (i.e., definitions of Aboriginal status by Aboriginal ancestry versus Aboriginal identity), and 7) Aboriginal identifiers used in Statistic Canada's surveys were not consistent with the Council of Ministers of Education, Canada (CMEC) Working Group on the Aboriginal Education Action Plan's recommendation of using "First Nations, Métis, and Inuit" (202).

3.3.2 Inadequate ethnicity/race information in administrative data

Health care administrative data are rich sources of demographic, health status, health service demands, and health service utilization information essential for public health research and applications. Administrative data is generally more cost- and time-effective than population

surveys since the collection of administrative records do not incur additional cost nor do they impose further burden on subjects. Unfortunately, the overwhelming majority of regional and provincial/territorial administrative databases in Canada, including the Canadian Cancer Registry (CCR), Discharge Abstract Database (DAD), National Physician Database (NPDB), Electronic Medical Records (EMR), and Health Insurance Registries (HIR), lack any ethnicity/race indicator. This is further demonstrated using a table excerpt from CIHI's pan-Canadian 2016 report (**Table 3-2**). Compared to age and sex, the administrative information on ethnicity/race and Aboriginal status data is either absent or inadequate in almost all CIHI data holdings, with respect to hospital and acute care, primary and physician care, drugs, disease and surgical registries, home and continuing care, mental health and rehab (**Table 3-2**). Only two (Canadian Patient Experiences Reporting System (CPERS) and Canadian Organ Replacement Register (CORR)) of the 14 databases contained both ethnicity/race and Aboriginal identity information. The categories in CPERS were White, Chinese, First Nations, Métis, Inuk or mixed, South Asian, Black, Filipino, Latin American, Southeast Asian, Arab, West Asian, Korean, Japanese, other, and unknown (203), whereas, the ethnic categories in CORR were Caucasian, Asian, Black, Indian sub-continent, Pacific Islander, Aboriginal, Mid East/Arabian, Latin American, other/multiracial, and unknown (204). Thus, not only are ethnicity, race, and Aboriginal status not frequently collected, when it exists, there may be a lack of comparability in response categories and category labelling.

3.3.3 Negative impacts from the inadequacy of ethnicity and race data

The general scarcity and inadequacy of ethnicity/race data in Canada hinder researchers' and decision-makers' ability to uncover, evaluate, and address health inequity issues related to disease burden and health care demand, quality, access, and utilization for the needed

ethnic/racial groups. For example, while many disease statistics stratified by ethnicity/race have been published in Canada in the past, the ethnicity/race data inadequacy has limited the scope, generalizability, and methodological soundness of these statistics. Our knowledge and attempts to understand the disease burden and incidence stratified by ethnic/racial groups for a wide range of acute and chronic diseases (i.e., asthma, influenza, cancer, cardiovascular diseases, diabetes, disability, mental illnesses,) are hindered. Researchers interested in finding ethnically/racially-specific disease statistics needed to conduct their own surveys, which were usually done within a relatively short and specific time frame and over a small sample from a particular geographic region due to study constraints. This led to reduced generalizability to a wider population, to a community under different context, or to a different time frame. Researchers attempting to study ethnically/racially-specific disease statistics in this non-supportive data environment not only need to be innovative, but also need to endorse unwanted methodological complications.

The study by Young *et al.* (2015) could demonstrate the point. They set forth to estimate colorectal cancer incidence in Aboriginals in Ontario despite the lack of ethnicity/race indicator in Ontario Cancer Registry. They used Canadian Censuses to dichotomize Dissemination Areas (DA) into "high-Aboriginal identity" and "low-Aboriginal identity" regions based on the proportion of individuals reporting "yes" to having Aboriginal identity, and then linked the regions with corresponding cancer statistics. As a result, stratified cancer incidence could only be estimated at the ecological level (205). This subsequently led to a number of methodological challenges, including possible confounding at individual and ecological levels, limitation due to ecological inference/interpretation, arbitrary cut-off (≥33% of the population self-identified as Aboriginal) of the "high-Aboriginal identity" areas, and lack of generalizability to the Métis and off-reserve Aboriginal population (205). These methodological complications resulting from

working around the absence of ethnicity/race data compromised the methodological soundness and usefulness of research results. If ethnicity/race information is part of the cancer case description, ethnically-/racially-specific cancer statistics will be much easier to calculate and will be at the individual level (instead of ecological level). Similarly, CIHI's pan-Canadian 2016 report of equity in health care has been primarily carried out using area-level data linkage since individual-level data with both health care and health inequity stratifiers (such as ethnicity/race and Aboriginal status) is generally unavailable (183). In addition to disease statistics, research intended to study inequity in access to health care (such as physician visit and proximity) and health service quality (such as safety, timeliness, effectiveness, cost-efficiency, cultural sensitivity, and patient-centeredness) by ethnicity or race are also likely to be deterred or compromised by the lack of adequate ethnicity/race data.

The lack and inadequacy of ethnicity/race data across Canada is concerning compared to countries such as the United States (U.S.), where ethnicity/race indicator exists in the majority of health surveys and administrative datasets (206). Moubarac (2013) conducted a comprehensive review of the use of ethnicity/race in research that addressed health disparities. This review included 280 studies published in high impact-factor journals in the domains of public health and epidemiology from 2009-2011. Only 1.1% of these studies were from Canada, compared to 84.0% from the U.S. and 5.7% from the U.K. (207). A plethora of researchers have reported about the relative scarcity and inadequacy of ethnicity data in Canada (89-94, 208, 209). Parsons (2005) attributed this to the "profound lack of routine collection of race/ethnicity relevant data in Canada". Randall (2007) argued that the insufficiency and inadequacy of ethnicity data hindered researchers' ability to accurately describe ethnic inequality and racial discrimination in health care. The availability of high-quality data on race, ethnicity, and other characteristics of

individuals receiving health care is critical to documenting inequities in health and health care. Without supportive ethnicity/race data to help describe, understand, and track patterns of disease burden/occurrence and medical needs and usage, finding a complete, effective, and long-lasting solution for our current inequitable health and social system is an impossible task.

3.4 Recommendations to Overcome Ethnicity/Race Data Obstacles

Canada needs to provide accurate and reliable ethnicity and race data at various regional and organization levels to establish a strong and necessary foundation for public health research, policy, and intervention. While researchers, healthcare professionals, and decision makers have long recognized the importance and needs for ethnicity/race data in Canada, there is no real movement and commitment to enhance ethnicity/race data collection systems. The remaining of this position paper provides a number of key recommendations of how our nation can implement local- and nation-wide ethnicity/race data collection in a more proactive and effective manner.

3.4.1 Ensure supports from key stakeholders

Increasing the scope of ethnicity/race data collection will affect various key stakeholder groups including the researcher, healthcare professional, community health planner, government/organization leader, policy-maker, patient, and general public. Stakeholders' support to the national increment of ethnicity/race data collection is paramount. Currently, evidence on the attitude and level of support of ethnicity/race data collection in Canada is mixed. While most studies from different Canadian regions have indicated that community leaders and healthcare professionals generally acknowledge the importance of ethnicity/race data collection and are in support of that (94), which are similar to studies from the U.S. (210) and U.K. (211), evidence on public perception and support are mixed suggesting the Canadian general public has some

reservation and concerns in disclosing their ethnicity/race information in various conditions (212). Their comfort levels vary by various factors. For example, the general public was found to be more willing and comfortable to disclose the ethnicity/race information to physicians in person than in written survey questionnaires (212). To help gain support from various stakeholders, it is important to deliver a clear message on why it is important and beneficial to collect ethnicity/race data. The message should also be tailored to the stakeholder group. For example, the benefits of having available high quality ethnicity/data for epidemiologic researchers will be different from that for the general public.

Understanding and addressing reasons for potential opposition from stakeholders are equally important. For example, concerns regarding data privacy and confidentiality by the general public in Canada are cited as barriers for the support of ethnicity/race data collection (213). Likewise, the identification of ethnicity/race for ethnic minorities (such as Aboriginal and visible minorities) have unjustly put them in grievous situations, including residential school and cultural destruction for Aboriginals in more distant past (214), and the racism and racial profiling by law enforcement for Aboriginal, Muslim, and Black Canadians in recent decades (214, 215). As a result, collecting sensitive information such as self-identification of ethnicity/race can create feelings of anxiety or distrust in some individuals (216). Development and adherence to ethical and legal guidelines for sensitive data (including but not exclusive to ethnicity/race) need to be in place to address these very real concerns. Individual privacy and confidentiality concerns could be addressed at the organizational or institutional level by applying various data protection protocols, including fully-voluntary informed consent, anonymity, de-identification, data encryption, password-protected access, controlled access, and secure data collection facility (217). On the other hand, systemic issues such as racism and racial profiling should be addressed

at various levels including assistance to victims, public education programs, consultation with policy makers and legislators, and intervention in court cases (215). This will likely help alleviate anxiety and worry when individuals disclosing ethnicity/race information.

3.4.2 Set clear objectives and agendas

To increase implementation and support of ethnicity/race data across Canada is a long and ambitious endeavor that requires clear objectives and strong collaboration between data collection and governance organizations. For example, CIHI's pan-Canadian 2016 report identified a number of short term and long term agendas. The short-term agendas included continuing stakeholder engagement and expanding to Aboriginal communities, patient groups, other pan-Canadian organizations, research community, senior leaders, and health regions; communicating a rationale for collecting health equity stratifiers that takes into account different needs for equity information at national, jurisdictional, regional, and care provision levels; collaborating with Statistics Canada and other partners to develop or implement a conceptual framework of health equity that can be applied to CIHI's Health System Performance Measurement Framework; solidifying the governance of equity data collection by determining who will be leading and collaborating (183). The long-term agendas included working with national health agencies (such as Statistics Canada, Public Health Agency of Canada, and CIHI), the provincial and territorial jurisdictions, and other custodians to develop data sharing agreements to enable equity measurement, while addressing legislation and concerns over privacy and confidentiality; undertaking knowledge translation activities to provide guidance on how to use equity data at different levels and how equity data can be beneficial to different stakeholders; and aligning effective communication strategies with current political priorities and interests of politicians and decision-makers (183). Different organizations or stakeholder groups

have their own capability, priority, and ethnicity/race data needs. Thus, they should devise their own set of objectives and agendas with respect to ethnicity/race data collection. We propose asking a number of potentially useful guiding questions as below:

- What are the added benefits of having ethnicity/race data in the organization and its clients, and how may the data be used?
- What barriers and opposition may arise from collecting ethnicity/race data in the organization, and how can they be addressed?
- What are the possibilities of collaboration and data sharing within and between local regions?
- How can these collaborations and data sharing possibilities be realized, communicated, and facilitated?
- What existing supports and constraints (i.e., legal, social, infrastructural, and cultural) are currently in place that are relevant to or need to be addressed for joint collaboration and data sharing?
- How does obtaining ethnicity/race data fit into the current culture and mission statements of the organization?
- Which team and department will lead and perform the data collection and management?
- Which locations will the data be gathered?
- What resources (i.e., human, time, and financial) will be available, needed, and acquired?
- What categories will be used to identify the ethnic/racial groups? Will these categorizations be compatible with other definitions in use in other databases within and outside the organization?
- Which format of acquiring ethnicity/race information will maximize response rates?

- What monitoring system and indicators will be used to measure the success and failure of the data collection and management processes?
- Are there policies, guidelines, and legislations that need to be reviewed and adhered to?
- Who will provide guidance/technical assistance to data collection, validation, and management?
- Who and how should the data be collected from?
- How frequently will the data be collected?
- How can the data collection be continued sustainably in the future?
- Is data quality evaluation needed after implementation, and how will it be evaluated?
- Is pilot testing needed?

3.4.3 Preparation to address data and implementation challenges

Canada should devise its own national measurement standards and criteria based on its context, experiences, and priorities. A number of guidelines for measuring ethnicity and race is listed below based on Mays *et al.* (2003) article (87) primarily, along with our own recommendations:

1) Joint discussions at regional and national levels should be carried out to clearly define the concepts of ethnicity and race, to specify the ethnicity/race data needs, and to identify working partners and explore collaboration and data sharing initiatives;
2) National guidelines regarding data collection and data sharing for ethnic/racial information should be developed via joint effort with representatives from all major stakeholder groups (i.e., health agencies, jurisdiction, community leaders, etc.);

3) National measurement standards with respect to definition, labelling, and categorization of ethnic and racial grouping should be developed to ensure comparability and consistency of ethnic/racial information across different datasets;
4) Develop ethnic categories that are mutually exclusive from other ethnic groups, flexible to allow multiple responses and absorb future responses not yet encountered, and consistent through time but with the consideration of the changing and subjective nature of ethnicity;
5) Include data collection on complex social variables for which ethnicity/race is often used as a proxy, such as social status, neighbourhood context, perceived discrimination, social cohesion, social capital, social support, types of occupation, employment, emotional well-being, and perceived life opportunities;
6) Use the method of self-reporting of ethnicity and race due to the subjective nature;
7) Ensure ethnic/racial allocation methods that are sound and culturally appropriate to visible minority and disadvantaged groups;
8) Facilitate collaborations between different stakeholders including end-users, researchers, state and local health departments, and collectors of federal data to enhance the quality of the data collected and to develop consistent, reliable, and valid policies to facilitate the existing research, training, and policy agendas;
9) Develop mechanisms for linking records across government data systems in mind; and
10) Consider collecting other indicators such as socioeconomic position, acculturation, and language which have been shown to correlate with ethnicity/race.

3.4.4 Learn from the success and failure in other countries

Compared to Canada, the U.S. is currently collecting ethnicity/race data more frequently and in more public health-related domains, such as health and disease registries, administrative databases, and EMR. However, the data collection system for ethnicity/race in the U.S. is not perfect and is facing its own challenges. Currently, their data challenges include 1) despite full coverage of the entire national population, the sample sizes are not usually sufficient to provide statistically reliable estimates of health and health care information for smaller racial and ethnic group, 2) most surveys do not have sufficient sample size large enough to be representative of all, or even most, of the individual jurisdiction, 3) missing information in provider-based surveys (i.e., clinics and hospital administrative survey), and 4) inconsistency in data collection and lack standardization in data question/answer categories in disease surveillance system data collection (206). The National Research Council (2004) in the U.S. developed a set of recommendations to tackle these data quality and collection issues which should be examined by the Canadian counterpart to foresee, prepare, and develop our own appropriate strategies to handle similar data challenges if they occur.

3.5 Conclusions

In this position paper, we have shown why ethnicity (including Aboriginal status) and race are much needed information in Canada, a country with vast cultural diversity. The current scarcity of good quality ethnicity/race data is systemic and deep-rooted. Such data inadequacy hinders researchers and decision-makers' ability to tackle major public health issues such as health inequities, particularly for the Aboriginals and ethnic minorities. Nation-wide increment of collection of good quality ethnicity/race data is paramount in building the foundation to improve the health of all Canadians. While this realization is not new to many parties, actionable plans need to be developed, put in place, and enforced in all levels of the government and

organization. Our recommendations cover various scopes and angles that are relevant in Canada, intended to guide organizations in planning, implementing, and sustaining the process of ethnicity and race data collection.

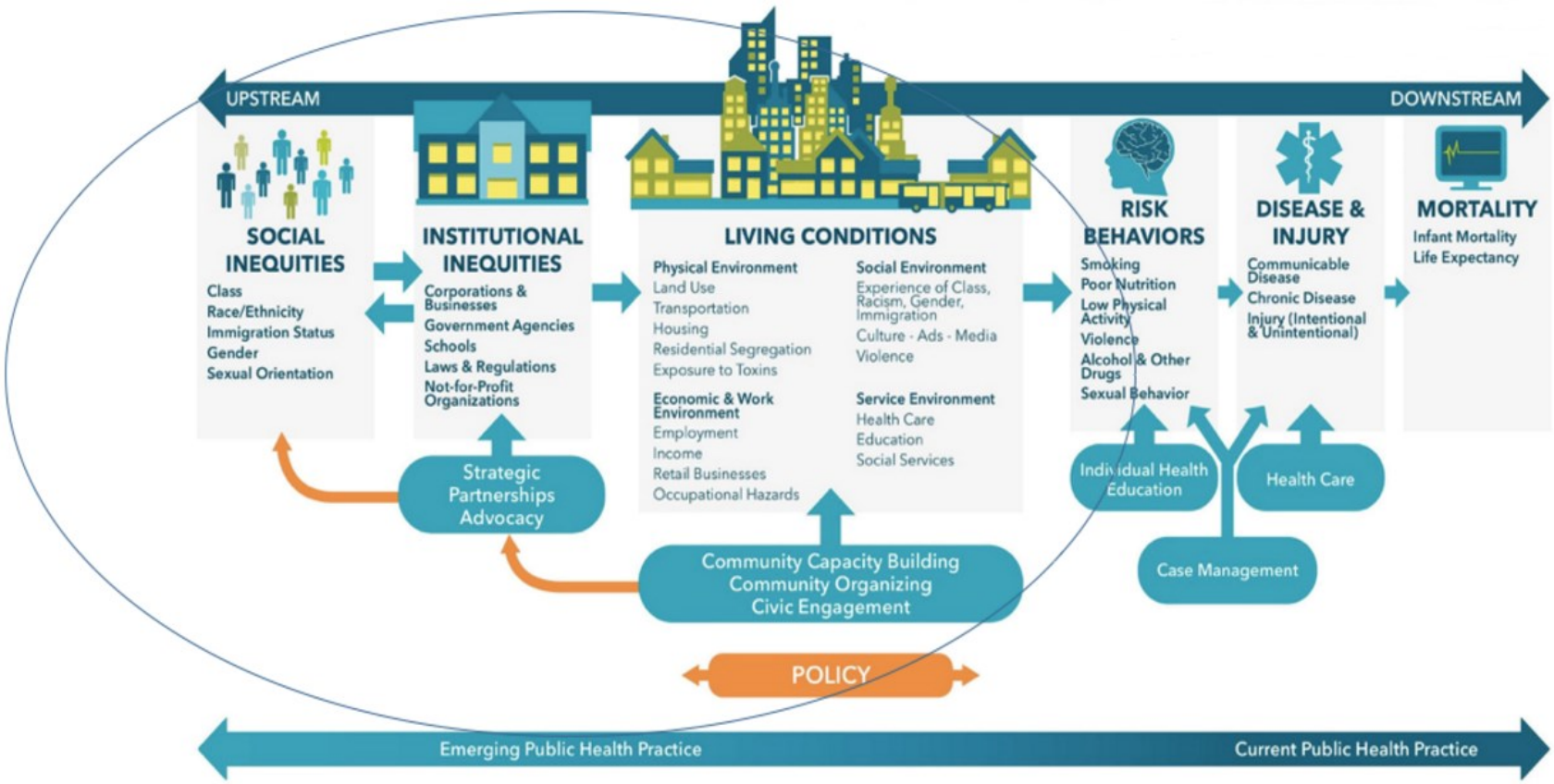


Figure 3-1: A framework for understanding the relationship between ethnicity/race and health outcomes (Extracted from (180)).

Table 3-1: Major active national surveys containing questions on ethnicity, race, or Aboriginal status in Canada.

	Coverage, Response rate, Frequency	Limitations
Canadian Census 2011 (218-220)	*Coverage*: Canadian population *Response rate*: Short form: 97.1%; long form (National Household Survey): 68.6% *Frequency*: Every five years	• Lacking extensive health-related questions. • Infrequent survey cycles. • Undercounting some groups such as homeless, young adults, and Aboriginals on reserves. • Changing wording of questions between censuses. • Unable to collect or present information from very small communities due to privacy and confidentiality principles. • Abandoning the mandatory long-form and replacing it by the voluntary National Household Survey 2011, which led to questionable data quality.
Canadian Community Health Survey 2012 (221-224)	*Coverage*: Random sampling of Canadian population aged 12 years and over	• Excluding children aged 11 years or younger and individuals/households without a telephone. • Excluding people living on reserves, certain Aboriginal settlements, and certain remote regions. • Certain questions (such as smoking and sexual behaviours) may be prone to errors from proxy reporting or social desirability bias when answering.

	Response rate: Combined household and person-level response rate: 68.9% for mental health component *Frequency*: Every year since 2007 (every two years prior to 2007)	• Lacking comparability for some question domains since different regions may choose different optional modules (i.e., physical check-up, mood disorder, and sexual behaviours). • Limited analysis in small regions/communities with small sample size. • Limited direct questions on disease severity. • Limited primary care questions and depth of questions regarding the appropriateness of service use and the quality of care in clinical settings.
Canadian Health Measures Survey 2014-2015 (225)	*Coverage*: Random sampling of Canadian population aged 3-79 years living in the ten provinces *Response rate*: Age and sex combined response rates: 51.4-56.3%	• Excluding people living in the three territories and people living on reserves and other Aboriginal settlements in the provinces. • Requiring large commitment by participants as they needed to complete personal interview at the household and physically visit a collection site where their physical, blood, and urine measures were taken. • Poor response rates. • Potentially-biased sampling due to small number of voluntary sample (n=5,700) and small number of designated collection sites (n=16) across Canada, as candidates living far away from these sites were unlikely to participate.

	Frequency: Every two years	
Aboriginal Peoples Survey 2012 (226)	*Coverage*: Random sampling of Aboriginal population aged six years or older living off-reserves across Canada *Response rate*: 76.0% *Frequency*: Every five years	• Infrequent survey cycles. • Excluding people living on reserves and settlements and in certain First Nations communities in Yukon and the Northwest Territories and children aged five years or younger. • The sample of Aboriginal Peoples Survey was selected from individuals reported to have an Aboriginal status from the National Household Survey 2011, which only had a 68.9% response rate. Compounding with the mediocre response rate of Aboriginal Peoples Survey itself (76.0%), selection bias is likely. • Limited health questions compared to major health surveys (i.e., CCHS).

Table 3-2: Equity stratifiers embedded at the individual level in CIHI data holdings*†.

Category	CIHI data source	Age	Sex	Ethnicity /race	Aboriginal identity
Hospital and acute care	• Discharge Abstract Database	A	A	N/A	S/A
	• Hospital Morbidity Database (Quebec only)	A	A	N/A	S/A
	• National Ambulatory Care Reporting System	A	A	N/A	S/A
	• Canadian Patient Experiences Reporting System	A	A	A	A
Primary And physician care	• Patient-Level Physician Billing Repository	A	A	N/A	N/A
	• Primary Health Care EMR Content Standards	A	A	N/A	N/A
Drugs	• National Prescription Drug Utilization Information System Database	A	A	N/A	S/A
Disease And surgical registries	• Canadian Organ Replacement Register	A	A	A	A
	• Ontario Trauma Registry - Comprehensive Data Set	A	A	N/A	N/A
	• Canadian Joint Replacement Registry	A	A	N/A	N/A

Home and continuing care	• Continuing Care Reporting System	A	A	N/A	S/A
	• Home Care Reporting System	A	A	N/A	A
Mental health and rehab	• National Rehabilitation Reporting System	A	A	N/A	A
	• Ontario Mental Health Reporting System	A	A	N/A	A

*Direct excerpt from Canadian Institute for Health Information (2016) (183). †Legend: A=Available. N/A=Not available. S/A=Somewhat available (i.e., data is incomplete, has high non-response or requires additional validation). EMR=Electronic medical record.

Chapter 4 – Name- and Location-based Aboriginal Ethnicity Classification

4.1 Background

An ethnic group is a population with members sharing a number of social and cultural components such as historical memory, cultural practice, national experience, world view, religion, and language (227). Ethnicity is important in shaping one's self-identity (84, 85) and lifestyle behaviours (81, 82, 228). It is related to but different from the concept of race. While race categorizes individuals by physical traits such as skin colour, hair texture, and facial features, ethnicity is self-identified based on a subjective sense of belonging towards an ethnic group (229).

Canada is one of the most ethnically diverse countries in the world (230), yet its existing data infrastructure supporting ethnicity information provision is considered inadequate by a number of researchers (89-94, 208, 209). The ethnicity data inadequacy was characterized by Parsons (2005) as a "profound lack of routine collection" of ethnicity data. Randall (2006) stated this data inadequacy has hindered researchers' ability to accurately describe ethnic inequality in health domains. A review by Moubarac (2013) compared different countries in terms of the volume of public health publications using ethnicity or race as part of the analysis (207). Only 1.1% of all included studies were from Canada, compared to 81.4% from the United States (U.S.).

One approach to alleviate the problem of ethnicity data inadequacy is to use widely collected variables to classify/predict individual's ethnicity. Readily available variables such as name and residence location have shown potential in classifying ethnicity status. Mateos (2007) conducted a systematic review and identified 13 studies that applied name-based ethnicity

classification methods. The sensitivity of these studies ranged between 0.67-0.95, specificity 0.80-1.00, positive predictive value 0.70-0.96, and negative predictive value 0.96-1.00 (96). Three of the 13 studies (published prior to 2006) were conducted in Canada (123-125). These studies were relatively old (published between 1993-1999) and examined only a small number of ethnic categories (i.e., Chinese only,(123, 124) and South Asian and Chinese (125)) using simple sets of name features (i.e., entire surname only (124, 125)). In Canada, the lack of ethnicity data does a disproportionate disservice to the disadvantaged Aboriginal populations by keeping the underlying health patterns, root causes, and health service needs hidden and unaddressed (100, 186, 231). The ability to predict Aboriginal ethnicity using existing data will be valuable and could serve as a foundation in improving the health of Aboriginal Canadians. Furthermore, recent studies outside of Canada have shown residence location to improve the accuracy in estimating ethnicity (127, 232, 233). However, since every country is different in ethnic composition, history, and measurement standards, whether name and location together can accurately predict ethnicity in the Canadian context is unknown.

To date, no study investigated the ability to predict Aboriginal identity using the existing name and location information with a large population dataset in Canada. The key focus of this study is to evaluate the performance of an automated machine learning (ML) approach to predict Aboriginal ethnicity amongst Canadian residents using name and residence location features.

4.2 Methods

4.2.1 Data source

The first national census of Canada was conducted in 1871 (234). We used the 1901 Canadian census to provide all the required individual-level data fields, including name,

residence location, and ethnicity (152). This particular census was chosen because it was made publicly accessible by multiple online sources (149-151) and the association between name and ethnicity should be relatively well-preserved in this older time period with the expected lower frequencies of interracial marriages and related name changes (153). Between the mid 18th century and late 19th century, European colonists imposed many policies and programs that forced Aboriginal assimilation to non-Aboriginal societal constructs and displaced Aboriginals from their lands (235, 236). The Indian Act of 1876 further enforced the Aboriginal assimilation by abandoning their traditional ways of life, including outlawing traditional infant naming ceremonies and reassigning Aboriginals with western names (237). The use of 1901 census is an attempt to achieve a balance between the maturity of Canadian census and preservation of traditional Aboriginal naming practices in the data. Furthermore, the 1901 census was only three to four generations old, thus its overall generalizability to modern time should still hold true.

4.2.2 Ethnicity labels

Records (0.57%) with missing name and/or ethnicity entries were removed. The ethnicity field of the 1901 census was a free-form text entry. Data cleaning and ethnicity recategorization were performed based on a set of criteria. Misspellings and alternative spellings were corrected and standardized, respectively. Four Aboriginal categorizations were considered: 1) Aboriginal (all inclusive); 2) First Nations (FN), Métis, and Inuit; 3) major Aboriginal language groups (Algonquian, Athapaskan, Iroquoian, Kootenay, Salish, Siouan, Tsimshian, and Wakashan); and 4) major Aboriginal tribes (Algonquin, Blackfoot, Cree, Iroquois, Micmac, Mohawk, and Ojibwa). The final 1901 census contained 5,031,794 individuals, of whom, 95,131 (1.9%) were Aboriginals. A total of 85,760 of the Aboriginals could be identified in one of the three major Aboriginal subgroups: 61,263 FN, 23,900 Métis, and 597 Inuit. Classifications on non-

Aboriginal ethnic groups (including Chinese, English, French, Irish, Italian, Japanese, Russian, and Scottish) have also been examined. However, since Aboriginal classification is the focus of this study, results for non-Aboriginal ethnicity classification are presented in **Supplementary Table 4-1** as supplementary information only.

4.2.3 Feature generation

The name variable from the census was processed. Irrelevant information such as number (i.e., 0-9), punctuation (i.e., ! and ?), single-letter initial (i.e., "W" in "Henry W. Anderson"), and title abbreviation (i.e., Mr., Mrs., and Dr.) were removed. Fifteen training features were derived from names, while one feature was derived from residence location at the city-, town-, or district-level. If the residence location information was not available at these three levels, province information was used. The name features were based on single characters, substrings of characters, entire name entities, and phonetic representation. To illustrate the process of name feature extraction, the example "Mrs. Kate C. Hart Jones" would first be cleaned and converted into lower-case "kate hart jones". Then individual name features were extracted as follows:

- Last name: "jones"
- 2-letter substrings of last name: "jo", "on", "ne", "es"
- 3-letter substrings of last name: "jon", "one", "nes"
- 4-letter substrings of last name: "jone", "ones"
- First name: "kate hart"
- 2-letter substrings of first name: "ka", "at", "te", "e(space)", "(space)h", "ha". "ar", "rt"
- 3-letter substrings of first name: "kat", "ate", "te(space)", "e(space)h", "(space)ha", "har", "art"

- 4-letter substrings of first name: “kate”, “ate(space)”, “e(space)ha”, “te(space)h”, “(space)har”, “hart”
- First letter of last name: “j”
- Last letter of last name: “s”
- First letter of first name: “k”
- Last letter of first name: “t”
- Number of name entity (based on “Kate C. Hart Jones”): 4
- Double-metaphones (based on “kate”, “hart”, “jones” separately): “KT”, “HRT”, “JNS”, “ANS”
- Average length of name entity (based on “kate hart jones”): value = total character count/number of name entity = 13/3 = 4

4.2.4 Primary analysis using regularized logistic regression

The primary analysis of this study was conducted using the regularized logistic regression (LG) classifier. Both full data (N=5,031,794) and randomly selected subsamples (N=500,000) were analyzed with LR classifiers. The rationale of having two sample sizes is to be able to reduce computational time with subsample and to be able to gauge if classification performance plateaus or improves by increasing the sample size to the maximum. The dataset was split 50:50 randomly into training and validation sets. One-versus-the-rest binary classification was conducted for each ethnic category (i.e., Aboriginal versus non-Aboriginal; First Nations versus non-First Nations; and Inuit versus non-Inuit).

To describe the regularized LR classifier, let n be the total number of samples, $x = (x1, x2..., xd)$ be a vector of d features, and $h_\theta(x) = P(y=\text{“Aboriginal”}|\theta, x)$ be the estimated

probability of a person predicted as "Aboriginal" for ethnicity *y* given *x* and θ, where θ = (θ_1,…, θ_d) is a weight vector for the corresponding features. The probability $h_\theta(x) = \frac{1}{1+e^{-\theta^T x}}$ is based on the sigmoid function and *y* is a Bernoulli random variable. The objective of LR is to minimize the following regularized cost function *J(θ)* (238, 239):

$$J(\theta) = \frac{-1}{n}\sum_{i=1}^{n} y_i log(h_\theta(x_i) + (1 - y_i)\, log\big(1 - h_\theta(x_i)\big) + \frac{\lambda}{2n}\sum_{j=1}^{d}\theta_j^2$$

The first term is the cost (negative log likelihood) function, which penalizes misclassification during training. The second term is the L2 regularization term which penalizes large elements of θ's and minimizes overfitting, where λ is the regularization parameter which serves to control the trade off between model fit to data and magnitudes of θ elements. The optimal λ value was determined indirectly via the specification of the penalty parameter *C* (in the Python Scikit-learn module) which is an inverse of regularization strength (i.e., smaller values mean stronger regularization) (240). The default value of 1 of parameter *C*, where $C = \frac{1}{\lambda}$, was chosen a-priori in our study. The minimization of *J(θ)* was done by gradient descent (238, 239). Gradient descent iteratively updates all θ's simultaneously until convergence and its rate of descent is controlled by a prespecified learning rate parameter.

4.2.5 Secondary analysis using support vector machines and decision trees

The secondary analysis was conducted using the support vector machines (SVM) and decision trees (DT) classifiers. The rationale of this analysis is to explore the performance of other machine learning algorithms. The secondary analysis is similar to the aforementioned primary analysis with LR except 1) no analysis was done using the full dataset and 2) grid search was used to obtain optimal hyperparameter values for DT classifiers. The hyperparameters, being

different from model parameters, are parameters that express higher-level properties of the model, such as how fast it will learn or its complexity/capacity to learn.

For SVM classifiers, the predicted *y* is a binary prediction {-1 (i.e., "non-Aboriginal"), +1 (i.e., "Aboriginal")}. The equation of the separating hyperplane is given $w^T x + b = 0$ where *w* is a weight vector. While *w* specifies the spatial orientation of the hyperplane, *b* specifies the distance the hyperplane is away from the origin in the Cartesian coordinate. Two support hyperplanes bound around the separating hyperplane in the same direction. The positive support hyperplane is defined as $w^T x + b = +1$ and negative support hyperplane as $w^T x + b = -1$. Samples above the positive support hyperplane are classified as +1, while samples below the negative support hyperplane are classified as -1. Samples lying on either support hyperplane are called the support vectors. The shortest distance between the two support hyperplanes is called the margin (M). The goal in SVM is to maximize M since this typically results in lower generalization error. To find M, let a negative support vector $x_{(1)}$ be $w^T x_{(1)} + b = -1$ and its closest point on the opposite support hyperplane be $w^T(x_{(1)} + Mw) + b = +1$. Combining the two equations and isolating M gives $M = \frac{2}{w^T w}$. The maximization task for M can be converted into a minimization task for mathematical convenience (241). With the slack variable ξ's as soft margin that accommodates for outliers, the optimization problem is as follows (242):

$$\min_{w,b,\xi} \frac{w^T w}{2} + K\sum_{i=1}^{n} \xi_i$$

$$s.t. \quad y_i(w^T x_i + b) \geq 1 - \xi_i,$$

$$\xi_i \geq 0, i = 1, 2, \ldots, n$$

The parameter *K* controls the penalty of having outliers falling within the margin. The smaller the *K*, the less penalty imposed to have points violating the margin constraint, while the larger

the *K*, the greater the penalty (243). We used the default value of 1.0 for parameter *K* a-priori. Soft-margin Lagrangian was used to solve for the aforementioned minimization objective with respective constraints (242).

For DT classifiers, the ID3 algorithm is used to build the structure of decision trees (244). The ID3 applies a greedy search algorithm by finding a feature value that splits a parent (tree) node into multiple child nodes resulting in the largest information gain (*IG*), at each node. *IG* is measured by the impurity measure, in our study the Gini index (v) (245):

$$v(s) = 1 - \sum_{j=1}^{c} p(j|s)^2$$

where *c* is the total number of classes (i.e., 2 for binary classification), and *p(j|s)* is the relative frequency of class *j* at node *s*. The *IG* is the net difference between the impurity value of parent node (before splitting) and the weighted average of the impurity values of child nodes (after splitting) (245):

$$IG = v(parent) - \sum_{i=1}^{k} \frac{m_i}{m} v(i)$$

where *v(parent)* is the Gini index at a given parent node, which is split into *k* child nodes, *m* is the total number of samples at the parent node, m_i is the number of samples in child node *i*, and *v(i)* is the Gini index at child node *i*. The nodes will keep splitting by finding the split with the largest *IG* at each step until a stopping rule (i.e., a node has 0 impurity (i.e., Gini index=0), all feature values are the same, or pre-pruning conditions) is met.

To obtain the optimal hyperparameter values for DT, a random subset of data (N=100,000) was used, which was independent from the prospective training and validation datasets for model parameter estimation. The space for grid search was as follows: 1) maximum

tree depth: 10, 20, 40, 80, 160, 200; 2) minimum sample to split: 2, 4, 6, 8, 10, 12, 14; and 3) minimum sample in a leaf: 1, 2, 4, 6, 8, 10, 12, 14. The optimization of hyperparameters was conducted using all 16 features to optimize the accuracy for classifying the Aboriginal (all inclusive) identity, based on the 10-fold cross validation method.

4.2.6 Class imbalance and performance indicators

Class imbalance refers to the case frequency of one class being disproportionately low (minority class) compared to the rest of the members (majority class). Class imbalance was expected in our study since some of the Aboriginal subgroups generally represented very small percentages of the total population. For example, less than 0.2% of Canadian population were Inuit (246). When severe class imbalance occurs, accuracy alone as performance indicator will likely be inadequate since a classifier can simply and blindly classify all instances into the majority class while obtaining high accuracy. Thus, we employed a large set of classification performance indicators including accuracy, area under the Receiver Operating Characteristics curve (ROC), sensitivity, specificity, PPV, and NPV.

PPV is the number of true positives divided by the sum of true positives and false positives (247). It is a measure of classifiers' exactness. High PPV indicates relatively low false positives, suggesting high certainty of a positive label. Sensitivity is the number of true positives divided by the sum of true positives and false negatives (247). Sensitivity is a measure of classifiers' completeness. High sensitivity indicates relatively low false negatives, suggesting the classifier is able to capture most of the truly positive cases correctly. The ROC curve is created by plotting the true positive rate (or sensitivity) against the false positive rate (or (1-specificity)), which indicates how the number of correctly classified positive cases varies with the number of incorrectly classified negative cases over different decision boundary thresholds. The ROC is

equal to the probability that a classifier will rank a randomly selected positive sample higher than a randomly selected negative one (assuming "positive" ranks higher than "negative") (248). This indicates that the ROC is not affected by the underlying class distribution, thus insensitive to class imbalance.

4.2.7 Demonstrative application

To illustrate how accurate our classifiers perform in the application of estimating population disease statistics, a simulation was conducted based on our obtained FN classification performance and hypothetical disease prevalence in a population. The hypothetical population has a sample size of 100,000, composed of 2.5% FN and 97.5% non-FN. Three hypothetical disease prevalence scenarios were set up such that for disease A, the true prevalence for FN was 0.001 (or 10 cases per 10,000 people in a specific time point) and non-FN was 0.010; for disease B, the true prevalence for FN was 0.010 and non-FN was 0.001; and for disease C, the true prevalence for FN was 0.020 and non-FN was 0.001. This setup covers disease prevalence differential in either direction and in relatively large magnitude.

While having high sensitivity and PPV is desirable for a classifier, in reality, these values are usually in a state of trade-off determined by the threshold values of decision boundary. For example, higher decision boundary will lead to higher PPV, as it becomes more certain that cases labelled as "positive" are truly positive. However, this will also likely increase the number of false negative since more positive cases will no longer be meeting the cut-off and mistaken as "negative", thus weakening the sensitivity. In this demonstration, a line plot will be created using sensitivity and PPV plotted against decision threshold. From this plot, we will manually choose three decision boundary thresholds (that give high sensitivity, high PPV, and the point where sensitivity and PPV curves meet) and apply them to estimate the predicted disease prevalence for

both FN and non-FN groups in the aforementioned disease prevalence scenarios. The predicted prevalence will be compared to the true prevalence and 95% CI to demonstrate 1) the degree of deviation on estimating disease prevalence from the true values after using our classifiers to predict individuals' FN identity, and 2) which decision boundary threshold gives the best estimation of disease prevalence compared to the true values.

4.3 Results

4.3.1 Primary analysis (LR classifier)

The complete breakdown of the frequency of each ethnic group in the 1901 census is described in **Supplementary Figure 4-1**. The primary analysis indicating classification performance on Aboriginal (all inclusive), FN, Métis, and Inuit using LR in validation set is shown in Table 1. Various comparisons in performance are available, such as 1) 16 features (with location) versus 15 features (without location), 2) different Aboriginal groupings, 3) full data versus subsampling (N=500,000), and 4) one random seed versus average of five random seeds. Overall, the use of the full dataset achieved the best performance for each of the Aboriginal classification. The results using only 500,000 subsample (column 2 in **Table 4-1**) were comparable to those derived from the average of five random seeds (column 3 in **Table 4-1**), suggesting that results based on 500,000 subsample were adequately robust in general. Using name and location features together, our LR classifiers did considerably well in classifying Aboriginal (all inclusive) and FN groups, with close to 100% accuracy, ROC, specificity, and NPV, and 65% sensitivity and >80% PPV when using the full dataset (column 1 in **Table 4-1**). The use of only 500,000 subsample lowered different performance indicators by varying degrees, such as a decrease of 1-2% in ROC, 14-15% in sensitivity, and 5% in PPV (column 2 in **Table 4-2**). The location feature was important in classifying Aboriginal (all inclusive) and FN groups,

as without it, the classification performance was markedly worse, such as a decrease of 8-14% in ROC, 34-36% in sensitivity, and 1-5% in PPV (columns 4 and 5 in **Table 4-1**).

For the classification of Métis identity, the LR classifier achieved nearly perfect (≥98%) ROC, accuracy, specificity and NPV, as well as a good (78%) PPV using full data (column 1 in **Table 4-1**). However, its performance in sensitivity was poor (<35%). This suggested that while individuals identified as "Métis" had relatively high certainty that they were truly Métis, many Métis were misclassified as "non- Métis". This tendency seemed to exist to a greater extent for the classification of Inuit (column 1 in **Table 4-1**), indicated by even lower sensitivity. Such poor performance in Inuit classification was not unexpected as training sample of Inuit is extremely limited (N_{train}<320) even using the full dataset.

Table 4-2 provides the performance on the Aboriginal subgroups based on major language and tribal groupings, with 16 (both name and location) features. Overall, the ROC and accuracy ranged between 91-100% in subsample and 99-100% in full sample across these groups. Sensitivity ranged from 23% (Athapaskan) to 65% (Algonquian) in Aboriginal language groups, and 33% (Algonquin) to 52% (Cree) in Aboriginal tribal groups using full data. PPV ranged from 72% (Wakashan) to 89% (Athapaskan) in Aboriginal language groups, and 70% (Cree) to 91% (Micmac) in Aboriginal tribal groups using full data. Our LR classifiers appeared to perform fairly well (>60% sensitivity and >75% PPV) in classifying the Algonquian and Kootenay groups (**Table 4-2**). However, the classification for most remaining sub-Aboriginal groups tended to have poor sensitivity, possibly due to severely limited available sample to train with.

4.3.2 Secondary analysis (SVM and DT classifiers)

Compared to **Table 4-1**, the LR and SVM classifiers performed comparably at 500,000 subsample, while DT classifier generally performed slightly worse than LR and SVM classifiers (**Table 4-3**).

Using the DT classifier, the top five most informative features (including and excluding the location feature) were identified (**Table 4-4**). For Aboriginal (all inclusive) and FN groups, the location features have been identified as the majority of the top five most informative features. The location features were less informative for Métis and Inuit classification, however, interpreting results for Métis and Inuit requires special caution due to the low number of available cases and generally poor sensitivity scores. When considering only the 15 name features, the number of name entity, average length of name entity, name substrings, entire first name, entire last name, first letter of first name, and double metaphones have been identified as top five most informative features (**Table 4-4**).

4.3.3 Demonstrative application

Using the FN classification results derived from the same analytical condition as in the second column of **Table 4-1**, the tradeoff between sensitivity and PPV via decision boundary values is illustrated in **Figure 4-1**. Three arbitrary decision boundary thresholds were chosen for our simulation: 1) 0.90 (sensitivity=0.21, PPV=0.87), 2) 0.10 (sensitivity=0.77, PPV=0.52), and 3) 0.23 (sensitivity=0.66, PPV=0.66).

The three pairs of the obtained sensitivity/PPV were used, along with the underlying hypothetical FN/non-FN composition, to calculate the predicted true positive (tp), true negative (tn), false positive (fp), and false negative (fn). By applying the true disease prevalence accordingly, the predicted disease cases and predicted disease prevalence were computed (full

calculations in **Supplementary Table 4-2**). The accuracy of the predicted disease prevalence was evaluated against the true disease prevalence and 95% C.I., as well as the absolute and relative differences in prevalence (**Table 4-5**). Overall, the best decision boundary threshold was 0.90 for the three disease scenarios. The predicted prevalence for FN were all within the 95% C.I. of true values at 0.90 decision boundary. Despite having low sensitivity, all except one (in Disease C scenario) of the predicted prevalence for non-FN were within the 95% C.I. of true values (**Table 4-5**). As the decision boundary decreased, so did PPV. While this resulted in minor improvement in the approximation of prevalence for non-FN, the accuracy of the predicted prevalence for FN suffered markedly. A reminder should be made that this demonstrative application was based on 500,000 subsample, thus, the predicted prevalence were expected to be superior given the full dataset.

4.4 Discussion

This study has provided positive evidence that using frequently-collected variables name and location can relatively accurately classify four Aboriginal groupings: Aboriginal (all inclusive), FN, Algonquian, and Kootenay. In general, the residence location is an important feature in classifying Aboriginal Canadians, which is likely due to certain geographic regions to be more populated by Aboriginal Canadians (249). Métis classification's performance was generally poor with respect to sensitivity. Difficulty in accurately predicting Métis identity was expected as they are, by definition, individuals with a mixed indigenous and European ancestry. Despite the use of the entire census data, many Aboriginal subgroups suffered from very low sample size, which was likely to be the primary reason for poor sensitivity performance.

The ability to identify Aboriginal individuals or specific subgroup of Aboriginal individuals in data using ML approach could potentially serve various important public health

functions including stratifying disease statistics by Aboriginal status, measuring residential segregation, monitoring migration, evaluating equal opportunities policies and political empowerment processes, identifying health needs, and improving health services to Aboriginal people (96, 97). Our demonstrative application illustrated that our classifiers could help produce predicted ethnically-specific disease statistics that are comparable to the underlying truth. The cancer registry is an example where the name/location-based Aboriginal ethnicity classification may be implemented. Limited evidence has shown cancer prevalence differed among First Nations, Métis, Inuit, and non-Aboriginal (250), yet almost all cancer registries in Canada do not collect Aboriginal or ethnic information. Another potential application of name-ethnicity classifier is to assist in the self-identification initiatives for Aboriginal students in various academic institutions (251, 252). This allows institutions to provide assistance that is more relevant to Aboriginal students by addressing academic needs, promoting institutional changes to meet the needs of Aboriginal learners and communities, promoting self-empowerment amongst Aboriginals' younger generations, and developing effective teaching methods that honor and suit Aboriginals' unique worldviews, cultures, and learning styles.

Personal name and location are recorded in many health and general administrative databases in Canada. The results shown in our study suggests that name/location-based ethnicity classification could potentially be a feasible, reliable, and economic solution to help fill the existing ethnicity data gap. This can, in turn, help identify and serve vulnerable populations such as Aboriginal Canadians as they generally experience greater social injustice and disadvantages in the healthcare, social, and political systems (100). The ability to quickly and accurately stratify health and disease patterns by Aboriginal status could benefit various stakeholders. Health researchers, government policy makers, and health care professionals will be equipped

with a viable tool to assess disease burdens, monitor community needs, and develop culturally-sensitive policies and health programs for specific Aboriginal people and communities (253).

Implementing name/location-based ethnicity classifiers such as ours in real-life public health settings requires end-user to determine how much sensitivity-PPV tradeoff is best suited for the specific project objective at hand. The demonstrative application indicated that our classifiers (even using only one-tenth of full data) appeared to lead to fairly accurate predicted disease prevalence in various disease prevalence conditions when using high decision boundary threshold that resulted in high PPV. Encouragingly, the high accuracy in the computed disease prevalence for the Aboriginal group does not largely and negatively affect the approximation of the disease prevalence for the non-Aboriginal group. This highly suggests our Aboriginal classifiers' (Aboriginal (all inclusive) and First Nations) usability in public health research and applications. Researchers and practitioners need to consider carefully of the roles and relationships between various parameters (such as disease prevalence, population size and differential between ethnic classes, and difference in importance of false positive versus false negative) to effectively decide if the name/location-based ethnicity classification approach is justifiable and what decision threshold value(s) to apply to maximize the likelihood of obtaining the most useful statistics.

4.5 Conclusion

Our study is one of the most comprehensive name/location-based ethnicity classification study for Aboriginal status in Canada. This study has produced and validated with a novel set of name features with and without the addition of residence location. Our results have shown positive evidence to support the value of such approach within the Canadian context. This study

established the ties between name, residence location, and Aboriginal ethnicity that might be effectively utilized to predict the widely missing Aboriginal status information in Canada.

Prospective studies should examine a more recent census or other data to validate the generalizability to more recent time. Multiple censuses might be combined to increase the number of training sample especially for the identified low-frequency Aboriginal subgroups. Widespread experimentation and implementation of this approach across Canada is improbable unless advocacy and awareness are initiated and pushed forward extensively, including the academic and non-academic health research, government, and clinical sectors. Consensus, standards, practice guidelines should also be developed, validated, and upheld to ensure accountability, privacy, and data quality and security for using this approach in the future.

Table 4-1: Performance measures of name- and location-based ethnicity classification using logistic regression classifier for Aboriginal and three major Aboriginal subgroups using 1901 Canadian census.*

	With location			Without location	
	1) LR (N=5MM+)	2) LR (N=500k)	3) LR (N=500k, average of 5 random seeds)	4) LR (N=5MM+)	5) LR (N=500k)
Aboriginal					
Validation sample	47,526	4,679	4,667	47,686	4,655
Accuracy	0.99	0.99	0.99	0.99	0.98
ROC	0.99	0.97	0.97	0.90	0.83
Sensitivity	0.65	0.51	0.52	0.31	0.15
Specificity	1.00	1.00	1.00	1.00	1.00
PPV	0.86	0.81	0.82	0.85	0.78
NPV	0.99	0.99	0.99	0.99	0.98
First Nations					
Validation sample	30,573	3,005	3,015	30,671	3,032
Accuracy	0.99	0.99	0.99	0.99	0.99
ROC	0.99	0.98	0.98	0.91	0.85
Sensitivity	0.65	0.50	0.50	0.31	0.16
Specificity	1.00	1.00	1.00	1.00	1.00
PPV	0.82	0.77	0.77	0.83	0.72
NPV	1.00	0.99	0.99	1.00	0.99
Métis					
Validation sample	12,017	1,226	1,201	11,966	1,176
Accuracy	1.00	1.00	1.00	1.00	1.00
ROC	0.98	0.94	0.94	0.89	0.75
Sensitivity	0.34	0.19	0.17	0.13	0.03
Specificity	1.00	1.00	1.00	1.00	1.00
PPV	0.78	0.65	0.66	0.77	0.55
NPV	1.00	1.00	1.00	1.00	1.00
Inuit					
Validation sample	301	20	24	314	26
Accuracy	1.00	1.00	1.00	1.00	1.00
ROC	1.00	0.98	0.99	0.93	0.85
Sensitivity	0.21	0.00	0.02	0.06	0.00
Specificity	1.00	1.00	1.00	1.00	1.00
PPV	0.63	0.00	0.20	0.66	0.00
NPV	1.00	1.00	1.00	1.00	1.00

*LR = logistic regression with l2 penalty parameter C at 1.0; N is the sum of training and validation sample (50:50) combined; k = 1,000 and MM = 1,000,000; “With location” = 15 name features and 1 location feature; “Without location” = 15 name features only; “validation sample” is the number of cases belonged to the corresponding ethnic group in validation data.

Table 4-2: Performance measures of ethnicity classification (using 15 name and 1 location features) for major Aboriginal language groups and Aboriginal tribes using 1901 Canadian census.*

	Major Aboriginal languages								Major Aboriginal tribes						
	Alg-an	**Iro-an**	**Ath**	**Wak**	**Sio**	**Sal**	**Tsi**	**Koo**	**Cre**	**Oji**	**Mic**	**Bla**	**Iro-is**	**Moh**	**Alg-in**
N = 500,000															
Validation sample	2,132	337	99	94	87	93	78	50	813	788	117	167	110	101	77
Accuracy	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
ROC	0.98	0.97	0.98	1.00	0.98	0.99	0.96	0.97	0.99	0.98	0.95	1.00	0.98	0.98	0.91
Sensitivity	0.48	0.28	0.01	0.07	0.06	0.15	0.04	0.18	0.33	0.24	0.09	0.34	0.14	0.33	0.05
Specificity	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
PPV	0.72	0.80	0.25	0.54	0.22	0.82	1.00	0.82	0.50	0.60	0.79	0.77	0.94	0.79	1.00
NPV	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
N = 5,031,794															
Validation sample	21,717	3,679	1,089	1,045	1,012	928	565	448	8,401	8,028	1,196	1,874	1,180	995	771
Accuracy	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
ROC	0.99	0.99	1.00	1.00	0.99	1.00	1.00	1.00	1.00	0.99	1.00	1.00	0.99	0.99	0.99
Sensitivity	0.65	0.53	0.23	0.37	0.30	0.53	0.33	0.63	0.52	0.45	0.42	0.48	0.36	0.51	0.33
Specificity	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
PPV	0.78	0.87	0.89	0.72	0.75	0.87	0.85	0.85	0.70	0.78	0.91	0.83	0.87	0.82	0.88
NPV	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

*ML classifier used was LR with l2 penalty parameter and C parameter at 1.0; N is the sum of training and validation sample (50:50) combined; "Validation sample" is the number of cases belonged to the corresponding ethnic group in validation data. Abbreviations: Alg-an=Algonquian, Iro-an=Iroquoian, Ath=Athapaskan, Wak=Wakashan, Sio=Siouan, Sal=Salish, Tsi=Tsimshian, Koo=Kootenay, Cre=Cree, Oji=Ojibwa, Mic=Micmac, Bla=Blackfoot, Iro-is=Iroquois, Moh=Mohawk, Alg-in=Algonquin.

Table 4-3: Performance measures of name- and location-based ethnicity classification using support vector machines and decision trees classifiers for Aboriginal and three major Aboriginal subgroups using 1901 Canadian census.*

	With location		Without location	
	2) SVM (N =500k)	**2) DT (N =500k)**	**2) SVM (N =500k)**	**2) DT (N =500k)**
Aboriginal				
Validation sample	4,710	4,710	4,736	4,742
Accuracy	0.99	0.99	0.98	0.98
ROC	0.97	0.83	0.80	0.55
Sensitivity	0.54	0.52	0.21	0.09
Specificity	1.00	1.00	1.00	1.00
PPV	0.75	0.79	0.57	0.66
NPV	0.99	0.99	0.98	0.98
First Nations				
Validation sample	2,981	3,022	3,050	2,984
Accuracy	0.99	0.99	0.99	0.99
ROC	0.97	0.78	0.82	0.56
Sensitivity	0.55	0.54	0.20	0.09
Specificity	1.00	1.00	1.00	1.00
PPV	0.72	0.73	0.56	0.65
NPV	0.99	1.00	0.99	0.99
Métis				
Validation sample	1,279	1,178	1,212	1,189
Accuracy	0.99	1.00	0.99	1.00
ROC	0.93	0.76	0.70	0.52
Sensitivity	0.26	0.19	0.09	0.04
Specificity	1.00	1.00	1.00	1.00
PPV	0.48	0.45	0.34	0.48
NPV	1.00	1.00	1.00	1.00
Inuit				
Validation sample	20	46	27	41
Accuracy	1.00	1.00	1.00	1.00
ROC	1.00	0.58	0.86	0.54
Sensitivity	0.05	0.11	0.00	0.00
Specificity	1.00	1.00	1.00	1.00
PPV	0.20	0.21	0.00	0.00
NPV	1.00	1.00	1.00	1.00

*SVM = support vector machines classifier with penalty parameter K at 1.0; DT = decision trees classifier with maximum tree depth at 40, minimum sample per leaf at 1, and minimum sample to split at 1; N is the sum of training and validation sample (50:50) combined; k = 1,000; "With location" = 15 name features and 1 location feature; "Without location" = 15 name features only; "validation sample" is the number of cases belonged to the corresponding ethnic group in validation data.

Table 4-4: Top five most informative features based on decision trees classifiers (N=500,000) for each Aboriginal group.*

		1 (most important)	2	3	4	5 (fifth most important)
Aboriginal (all inclusive)	*With location*	location= unorganized territories/territoires non-organisés, the territories, Canada	location= unspecified, Alberta, the territories, Canada	location= unspecified, Assiniboia (east/est), the territories, Canada	location= Saskatchewan, the territories, Canada	location= unspecified, Algoma, Ontario, Canada
	Without location	name-entity	avg-length	lastLetter-lastName='k'	substring='ke'	substring='chel'
First Nations	*With location*	location= unorganized territories/territoires non-organisés, the territories, Canada	location= unspecified, Alberta, the Territories, Canada	location= unspecified, Assiniboia (east/est), the territories, Canada	location= unspecified, Algoma, Ontario, Canada	avg-length
	Without location	name-entity	avg-length	substring='as'	substring='ks'	substring='ah'
Métis	*With location*	location= Saskatchewan, the territories, Canada	location= Selkirk, Manitoba, Canada	location= unorganized territories/territoires non-organisés, the territories, Canada	location= unspecified, Algoma, Ontario, Canada	last-name=' peletier'
	Without location	metaphone='KRTNL'	metaphone='FTLR'	metaphone='PRNKN'	substring='jarl'	last-name='pruden'
Inuit	*With location*	substring='okta'	first-name='apalluk'	first-name='taursecak'	metaphone='AKTK'	last-name='kooitark'
	Without location	metaphone='PNKSL'	metaphone='APLK'	last-name='kooitark'	substring='kayr'	last-name='udgarder'

*Decision trees classifier with maximum tree depth at 40, minimum sample per leaf at 1, and minimum sample to split at 1; Total N = 500,000 (50:50 split to training and validation sets); "With location" = 15 name features and 1 location feature; "Without location" = 15 name features only.

Table 4-5: The comparisons of predicted and true disease prevalence in various disease scenarios and decision boundary thresholds.*

	Disease A		Disease B		Disease C	
	FN	**Non-FN**	**FN**	**Non-FN**	**FN**	**Non-FN**
True prevalence (95% C.I.)	0.0010 (0.0002-0.0029)	0.0100 (0.0094-0.0106)	0.0100 (0.0065-0.0147)	0.0010 (0.0008-0.0012)	0.0200 (0.0149-0.0263)	0.0010 (0.0008-0.0012)
Decision boundary=0.90 (sensitivity=0.21; PPV=0.87)						
Predicted prevalence	**0.0022**	**0.0098**	**0.0088**	**0.0012**	**0.0175**	0.0014
Prevalence difference	0.0012	-0.0002	-0.0012	0.0002	-0.0025	0.0004
Prevalence ratio	2.20	0.98	0.88	1.20	0.88	1.40
Decision boundary=0.23 (sensitivity=0.66; PPV=0.66)						
Predicted prevalence	0.0041	**0.0099**	**0.0069**	**0.0011**	0.0135	**0.0012**
Prevalence difference	0.0031	-0.0001	-0.0031	0.0001	-0.0065	0.0002
Prevalence ratio	4.10	0.99	0.69	1.10	0.68	1.20
Decision boundary=0.10 (sensitivity=0.77; PPV=0.52)						
Predicted prevalence	0.0053	**0.0099**	0.0057	**0.0011**	0.0109	**0.0011**
Prevalence difference	0.0043	-0.0001	-0.0043	0.0001	-0.0091	0.0001
Prevalence ratio	5.30	0.99	0.57	1.10	0.55	1.10

*FN = First Nations, non-FN = non-First Nations. Prevalence difference = predicted prevalence – true prevalence. Prevalence ratio $= \frac{\text{predicted prevalence}}{\text{true prevalence}}$. Bolded figures are predicted prevalence that fall within the range of the 95% C.I. (calculated using binomial "exact" method) of the true prevalence.

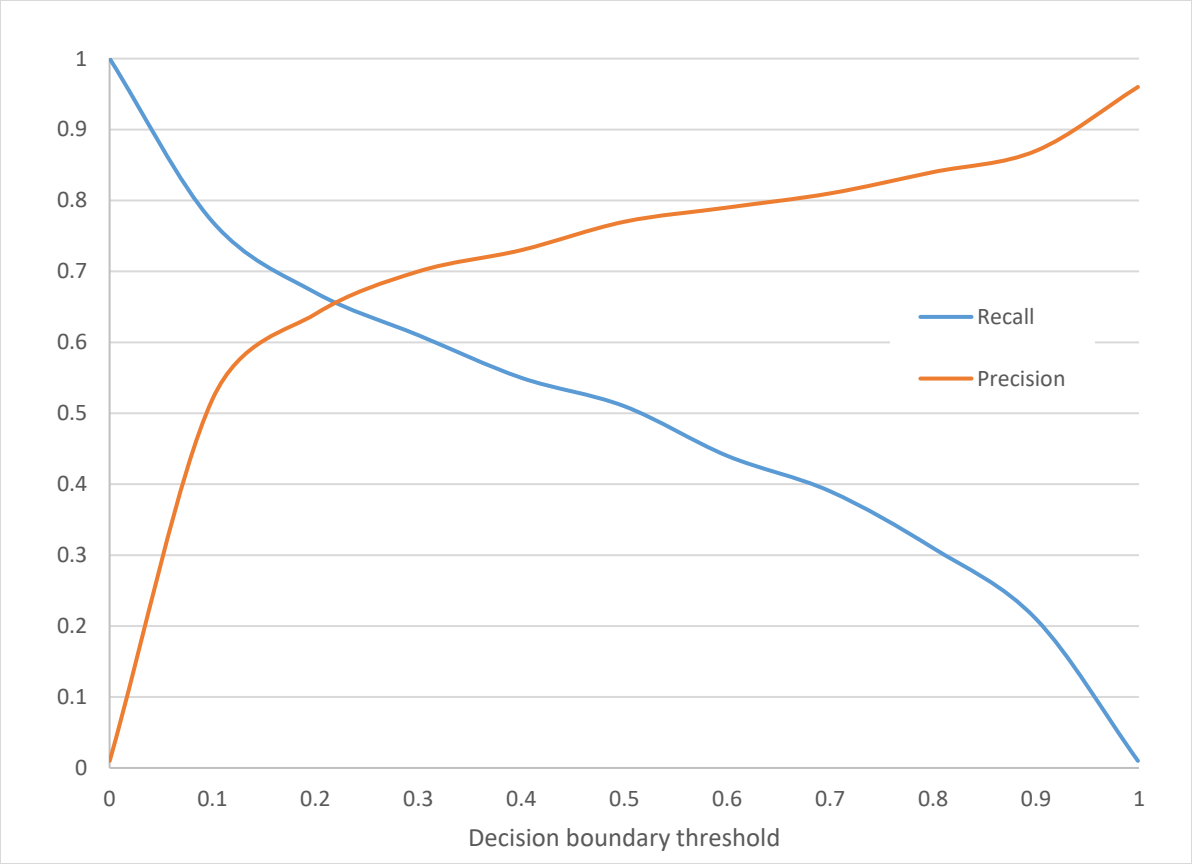
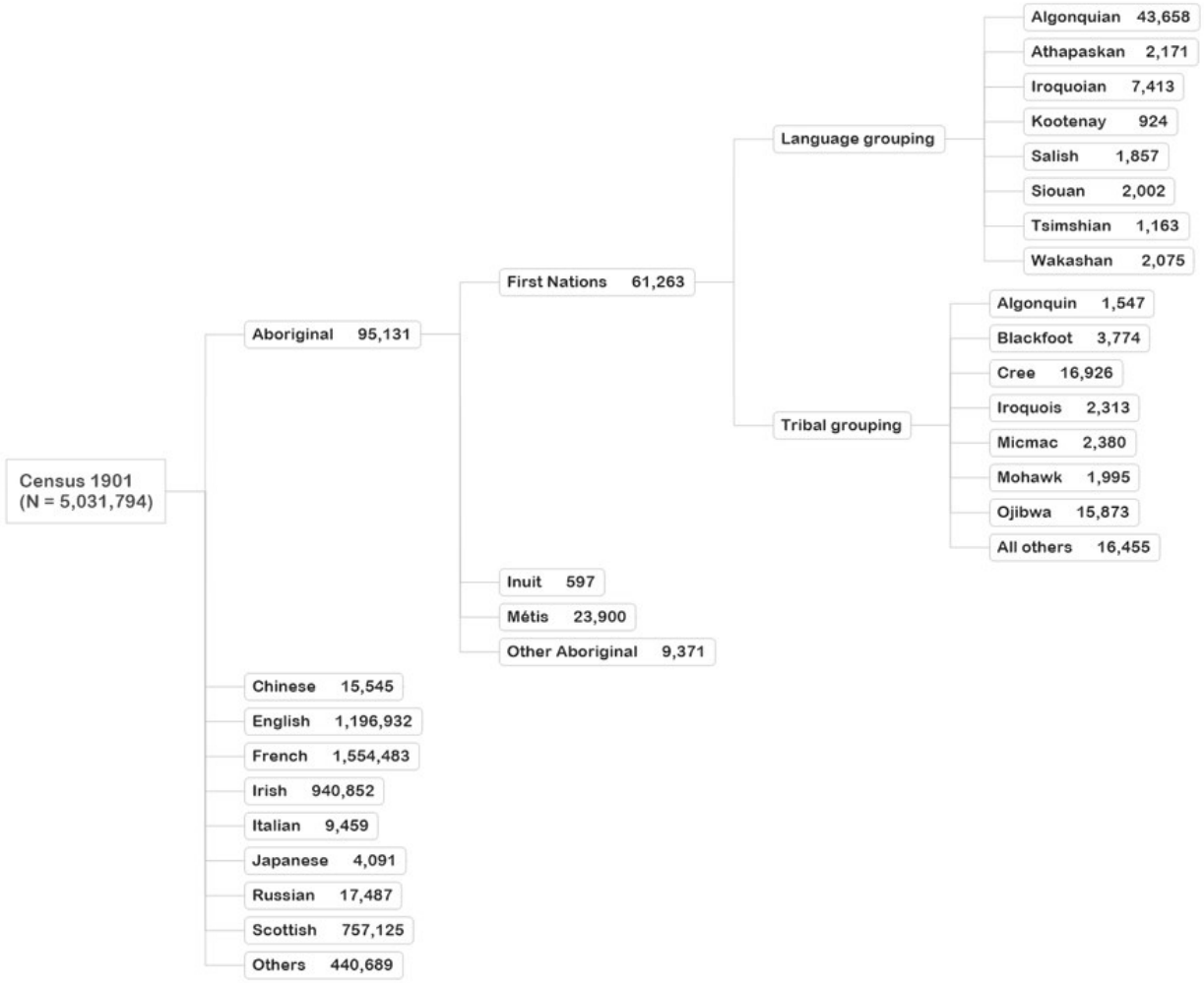


Figure 4-1: Sensitivity-PPV tradeoff for First Nations classification (logistic regression classifier, N=500k, all 16 features).



Supplementary Figure 4-1: Frequency of ethnic groups in 1901 Canadian census.

Supplementary Table 4-1: Performance measures of name- and location-based ethnicity classification for non-Aboriginal ethnic groups using 1901 Canadian census.*

	With location			Without location	
	1) LR (N=5MM+)	**2) LR (N=500k)**	**3) LR (N=500k, average of 5 seeds)**	**4) LR (N=5MM+)**	**5) LR (N=500k)**
Chinese (validation sample: ~798 at 500k and ~7,787 at 5MM)					
Accuracy	1.00	1.00	1.00	1.00	1.00
ROC	1.00	1.00	0.99	1.00	0.99
Sensitivity	0.91	0.86	0.87	0.88	0.83
Specificity	1.00	1.00	1.00	1.00	1.00
PPV	0.96	0.95	0.95	0.95	0.95
NPV	1.00	1.00	1.00	1.00	1.00
English (validation sample: ~59,536 at 500k and ~598,195 at 5MM)					
Accuracy	0.85	0.83	0.83	0.84	0.82
ROC	0.90	0.88	0.88	0.89	0.86
Sensitivity	0.61	0.55	0.55	0.57	0.51
Specificity	0.92	0.91	0.91	0.92	0.91
PPV	0.71	0.67	0.66	0.69	0.64
NPV	0.88	0.87	0.87	0.87	0.86
French (validation sample: ~77,490 at 500k and ~776,813 at 5MM)					
Accuracy	0.97	0.96	0.96	0.96	0.95
ROC	0.99	0.99	0.99	0.98	0.98
Sensitivity	0.94	0.92	0.93	0.92	0.90
Specificity	0.98	0.98	0.98	0.98	0.97
PPV	0.96	0.95	0.95	0.95	0.94
NPV	0.97	0.97	0.97	0.97	0.96
Irish (validation sample: ~46,724 at 500k and ~470,326 at 5MM)					
Accuracy	0.87	0.86	0.86	0.87	0.86
ROC	0.90	0.88	0.88	0.89	0.86
Sensitivity	0.51	0.46	0.46	0.46	0.41
Specificity	0.96	0.95	0.95	0.96	0.96
PPV	0.74	0.69	0.70	0.73	0.69
NPV	0.89	0.89	0.89	0.89	0.88
Italian (validation sample: ~460 at 500k and ~4,709 at 5MM)					
Accuracy	1.00	1.00	1.00	1.00	1.00
ROC	0.98	0.94	0.95	0.96	0.92
Sensitivity	0.45	0.25	0.20	0.36	0.14
Specificity	1.00	1.00	1.00	1.00	1.00
PPV	0.86	0.85	0.76	0.83	0.71
NPV	1.00	1.00	1.00	1.00	1.00
Japanese (validation sample: ~191 at 500k and ~2,050 at 5MM)					
Accuracy	1.00	1.00	1.00	1.00	1.00
ROC	1.00	1.00	0.99	0.99	0.99
Sensitivity	0.78	0.57	0.58	0.64	0.46
Specificity	1.00	1.00	1.00	1.00	1.00
PPV	0.93	0.86	0.92	0.89	0.90

NPV	1.00	1.00	1.00	1.00	1.00
Russian (validation sample: ~890 at 500k and ~8,753 at 5MM)					
Accuracy	1.00	1.00	1.00	1.00	1.00
ROC	0.99	0.97	0.97	0.98	0.94
Sensitivity	0.63	0.36	0.33	0.55	0.23
Specificity	1.00	1.00	1.00	1.00	1.00
PPV	0.91	0.77	0.77	0.90	0.75
NPV	1.00	1.00	1.00	1.00	1.00
Scottish (validation sample: ~37,435 at 500k and ~378,400 at 5MM)					
Accuracy	0.91	0.90	0.90	0.90	0.90
ROC	0.92	0.90	0.90	0.91	0.89
Sensitivity	0.56	0.52	0.52	0.53	0.49
Specificity	0.97	0.97	0.97	0.97	0.97
PPV	0.77	0.75	0.75	0.76	0.74
NPV	0.93	0.92	0.92	0.92	0.92

*LR = logistic regression with l2 penalty parameter C at 1.0; N is the sum of training and validation sample (50:50) combined; k = 1,000 and MM = 1,000,000; "With location" = 15 name features and 1 location feature; "Without location" = 15 name features; "validation sample" is the number of cases belonged to the corresponding ethnic group in validation data.

Supplementary Table 4-2: Calculations of predicted prevalence in demonstrative application.[*]

Disease A		Disease B		Disease C	
	N (sample)		N (sample)		N (sample)
FN	2,500	**FN**	2,500	**FN**	2,500
non-FN	97,500	**non-FN**	97,500	**non-FN**	97,500
	Disease prevalence		Disease prevalence		Disease prevalence
FN	0.0010	**FN**	0.0100	**FN**	0.0200
non-FN	0.0100	**non-FN**	0.0010	**non-FN**	0.0010
	Disease N		Disease N		Disease N
FN	2.5	**FN**	25.0	**FN**	50.0
non-FN	975.0	**non-FN**	97.5	**non-FN**	97.5
Decision threshold = 0.90 (PPV = 0.87, sensitivity = 0.21)					
	Predicted N		Predicted N		Predicted N
FN(tp)	525.0	**FN**(tp)	525.0	**FN**(tp)	525.0
FN(fn)	1,975.0	**FN**(fn)	1,975.0	**FN**(fn)	1,975.0
non-FN(tn)	97,421.6	**non-FN**(tn)	97,421.6	**non-FN**(tn)	97,421.6
non-FN(fp)	78.4	**non-FN**(fp)	78.4	**non-FN**(fp)	78.4
	Predicted disease N		Predicted disease N		Predicted disease N
FN (tp+fp)	1.3	**FN** (tp+fp)	5.3	**FN** (tp+fp)	10.6
non-FN (tn+fn)	976.2	**non-FN** (tn+fn)	117.2	**non-FN** (tn+fn)	136.9
	Predicted prevalence		Predicted prevalence		Predicted prevalence
FN	0.0022	**FN**	0.0088	**FN**	0.0175
non-FN (tn+fn)	0.0098	**non-FN** (tn+fn)	0.0012	**non-FN** (tn+fn)	0.0014
Decision threshold = 0.10 (PPV = 0.52, sensitivity = 0.77)					
	Predicted N		Predicted N		Predicted N
FN(tp)	1,925.0	**FN**(tp)	1,925.0	**FN**(tp)	1,925.0
FN(fn)	575.0	**FN**(fn)	575.0	**FN**(fn)	575.0
non-FN(tn)	95,723.1	**non-FN**(tn)	95,723.1	**non-FN**(tn)	95,723.1
non-FN(fp)	1,776.9	**non-FN**(fp)	1,776.9	**non-FN**(fp)	1,776.9
	Predicted disease N		Predicted disease N		Predicted disease N
FN (tp+fp)	19.7	**FN** (tp+fp)	21.0	**FN** (tp+fp)	40.3

non-FN (tn+fn)	957.8	**non-FN** (tn+fn)	101.5	**non-FN** (tn+fn)	107.2
	Predicted prevalence		Predicted prevalence		Predicted prevalence
FN (tp+fp)	0.0053	**FN** (tp+fp)	0.0057	**FN (tp+fp)**	0.0109
non-FN (tn+fn)	0.0099	**non-FN** (tn+fn)	0.0011	**non-FN** (tn+fn)	0.0011
Decision threshold = 0.23 (PPV = 0.66, sensitivity = 0.66)					
	Predicted N		Predicted N		Predicted N
FN(tp)	1,650.0	**FN**(tp)	1,650.0	**FN**(tp)	1,650.0
FN(fn)	850.0	**FN**(fn)	850.0	**FN**(fn)	850.0
non-FN(tn)	96,650.0	**non-FN**(tn)	96,650.0	**non-FN**(tn)	96,650.0
non-FN(fp)	850.0	**non-FN**(fp)	850.0	**non-FN**(fp)	850.0
	Predicted disease N		Predicted disease N		Predicted disease N
FN (tp+fp)	10.2	**FN** (tp+fp)	17.4	**FN** (tp+fp)	33.9
non-FN (tn+fn)	967.4	**non-FN** (tn+fn)	105.2	**non-FN** (tn+fn)	113.7
	Predicted prevalence		Predicted prevalence		Predicted prevalence
FN (tp+fp)	0.0041	**FN** (tp+fp)	0.0069	**FN** (tp+fp)	0.0135
non-FN (tn+fn)	0.0099	**non-FN** (tn+fn)	0.0011	**non-FN** (tn+fn)	0.0012

FN = First Nations; non-FN = non-First Nations; tp = true positive = sensitivity(N of FN); fn = false negative = (N of FN) – tp; tn = true negative = (N of non-FN) – fp; fp = false positive $= \frac{\text{tp} - (\text{PPV} * \text{tp})}{\text{PPV}}$. Predicted disease N for FN = (tp*(true prevalence for FN))+(fp*(true prevalence for non-FN)); predicted disease N for non-FN = (tn*(true prevalence for non-FN))+(fn*(true prevalence for FN)). Predicted prevalence for FN $= \frac{\text{Predicted disease N for FN}}{\text{tp+fp}}$; predicted prevalence for non-FN $= \frac{\text{Predicted disease N for non-FN}}{\text{tn+fn}}$.

Chapter 5 – Sentiment Analysis of Breast Cancer Screening in the United States Using Twitter

5.1 Background

Breast cancer is the most prevalent cancer among women in the United States (U.S.) (254). Regular breast cancer screening is important in detecting breast tumors early. Screening mammogram, clinical breast exam (CBE) performed by health professionals, breast self-exam, and breast magnetic resonance imaging (MRI) are examples of breast screening tests. A systematic review by (255) concluded that among women with average risk (i.e., no personal or family history of breast tumor/lesion, or genetic mutations such as BReast CAncer susceptibility gene- (BRCA-)1 and BRCA-2), mammography was associated with 20% reduction in breast cancer mortality. The American College of Obstetricians and Gynecologists (ACOG) (2011) guidelines (115) recommended U.S. women aged 40-74 with average risk to attend a screening mammogram and CBE annually. Women aged 75 and above with average risk should consult with physicians to decide whether to continue receiving mammogram.

Not all U.S. women adhered to the recommended breast screening guidelines. The uptake of breast screening varied across residence location (116), social class (117), and ethnicity (118). Whether or not to seek breast screening often depended on one's perception regarding the quality of care, competency of health professionals, discomfort level during the procedure, and length of time waiting for the procedure and test results (120). Women not attending regular breast screening listed the main reasons as being busy, unaware of breast cancer risk, fearful of receiving a true cancer diagnosis or a false diagnosis, and deterred by the pain and discomfort from the procedure (121). Many of these reasons can be explained by the health belief model (HBM) (256) which states that individuals' readiness and commitment to adopt a new healthy

behaviour are built on four perception-based constructs, including perceived susceptibility, perceived severity, perceived benefits, and perceived barriers. Evidently individual's subjective perception formed about breast screening, including face-to-face physician recommendation and perceived effectiveness and safety of breast screening (257-259), played a crucial role in determining if a woman would attend regular breast screening. However, continuous and unfiltered perception data on medical procedures is unavailable in public health surveillance, administrative, and other health-related databases (260).

Twitter is a rich data source of perception data. Twitter is used by hundreds of millions of active users continuously broadcasting their uncensored opinions, experiences, thoughts, and feelings in a form of a tweet, a short text message of 140 characters or less (104, 106). A considerable portion of tweets was health-related (107, 108) and could contribute to various health monitoring applications such as public awareness of influenza (109), worldwide influenza incidence (110), self-reported mental illnesses (111), medical complaints (112), and safety events by hospital patients (113). As for cancer communities, Twitter served as a popular digital platform to bring together different groups of key stakeholders. Medical professionals used Twitter to disseminate scientific findings and connect with patients (143), cancer patients used it to share experience, gain support, and educate one another (144, 145), and general public used it to advocate and raise funding (146). No study was found to examine Twitter's potential in gauging public perception on preventive public health interventions such as breast cancer screening.

Sentiment analysis is a sub-domain of natural language processing and computational linguistics that extracts subjective information from a text and assigns a sentiment score or a sentiment polarity classification (i.e., neutral, positive, and negative) (261). Sentiment analysis

helps determine the attitude or perception of a writer with respect to a specific topic in a systematic and quantifiable manner. We propose a sentiment analysis that not only demonstrates the visualization of sentiment patterns using breast screening tweets in the U.S. (descriptive analysis), but also explores the relationship between breast screening sentiment from Twitter and actual breast screening uptake behaviour derived from an external data source (hypothesis-based analysis).

5.2 Methods

5.2.1 Breast screening tweets and tweet processing

Twitter allowed public access to 1% random subset of tweets via Twitter REST API (262). Via the API, tweets related to breast cancer screening published from 17th September 2014 to 10th May 2015 were collected using the following filtering terms, which was compiled based on literature review and reviewed by our thesis committee including cancer epidemiology experts:

> *"mammogram", "mammography", "breast imaging", "breast screening", "breast mri", "breast ultrasound", "breast self-exam", "breast examination", "breast exam", "breastselfexam", "breastexam"*

Extracted information from each breast screening tweet included user name, time of tweet, published tweet content, and two types of geographic information including user-described location and user-enabled GPS location in longitude and latitude (165).

The content of each tweet was processed by removing any retweet tag ("RT"), hashtag symbol ("#"), user-mention tag ("@"), and Uniform Resource Location (URL) links. Not all Twitter users have described location information or enabled the GPS option. If both location

inputs were available, the more precise GPS location was used, otherwise the user-described location was used. If existed, user-described location was converted into GPS coordinates using Python module Geocoder (by accessing MapQuest) (166). The location information was then standardized by reverse-geocoding the coordinates into corresponding country, state, county, and city.

5.2.2 VADER sentiment classifier

There are a number of existing automated sentiment classifiers (169), such as Linguistic Inquiry and Word Count (LIWC), General Inquirer (GI), Affective Norms for English Words (ANEW), SentiWordNet (SWN), SenticNet (SCN), Word-Sense Disambiguation (WSD), and Hu-Liu-2004. These sentiment classifiers were not developed specifically for microblogging platforms such as Twitter. Tweets generally employed unique communication patterns (i.e., hashtag, user-mention, all-capitalization, acronyms, emoticons, slangs, and repeated punctuations) to better express emotions and fit in the microblogging culture. (169) developed and made publically available a sentiment classifier, called Valence Aware Dictionary for sEntiment Reasoning (VADER) classifier, specifically tailored to microblogging platforms such as Twitter. The sentiment lexicon of VADER classifier was based on well-established and human-validated sentiment lexicons (i.e., from LIWC, GI, and ANEW) and extended by adding common microblogging vernaculars (i.e., acronyms, slangs, and emoticons). In addition, grammatical and syntactical aspects of text (i.e., use of repeated punctuation such as “!!!!” and all-cap such as “EXTREMELY GOOD day”) were incorporated by systematically adjusting the baseline sentiment value using a rule-based model (169).

To classify the sentiment of a text, the VADER classifier examines the sentiment polarity and intensity of each word of the text against its lexicon, and then outputs four VADER

sentiment scores: neutral, positive, negative, and composite scores. The neutral, positive, and negative scores correspond to the proportion of text containing a particular sentiment polarity. For example, a 1.0 positive sentiment score indicates that every word in a text contains positive sentiment while 0.0 positive score indicates zero positive word, and likewise for neutral and negative sentiment scores. The composite score is computed by summing the sentiment intensity score of each word from the text that has a match with the VADER lexicon, adjusted with grammatical and syntactical rules, and then normalized to be between -1 (most negative) and +1 (most positive). The composite score can be used as a single unidimensional measure of sentiment. (169) concluded the VADER classifier considerably outperformed seven established sentiment classifiers (i.e., LIWC, GI, and ANEW). The VADER classifier achieved a 0.99 precision, 0.94 recall, and 0.96 F1 score, which were comparable to human accuracy.

5.2.3 Modifications and implementation of VADER sentiment classifier

Although VADER was validated on general tweets by (169), its accuracy performance to classify sentiment of tweets related to public health intervention, specifically breast cancer screening, topic required further validation. Such validation was conducted in our study by drawing a random subset of 250 tweets from the original breast screening tweets pool. The composite score was categorized into neutral (-0.3 to +0.3), positive (>+0.3 to +1.0), and negative (-1.0 to <-0.3). The sentiment polarity (neutral, positive, and negative) of each of the 250 tweets was determined by a blind-rater KW as gold standard. A poor accuracy (<40.0%) was observed from the VADER classification initially and the primary reason was identified. In the original VADER lexical dictionary, the lexicon “cancer” contained a highly negative sentiment value (-3.4). This resulted in VADER universally assigned highly negative composite sentiment score to virtually all tweets since they were related to breast cancer by default. Similarly, other

words including “die”, “died”, and “death” containing highly negative default sentiment values (i.e., -2.9, -2.6, and -2.9, respectively) were identified, yet these lexicons often appeared in our collected tweets as part of the breast cancer statistics conversations without any default positive or negative connotation. The effect on sentiment classification accuracy was examined by removing these lexicons from the original lexical dictionary, resulting in more favorable accuracy (77.2%). The remaining classification discrepancy between VADER and human rater was derived from more advanced sentiment classification challenges such as sarcasm, sentiment ambiguity, and mixed sentiments that were difficult for even human raters and thus unlikely to be addressed by further minor modifications in the VADER classifier. This modified version of VADER classifier was used to compute sentiment scores of breast screening tweets.

5.2.4 Descriptive sentiment analysis

Temporal, geospatial, and thematic patterns of sentiment from breast screening tweets were examined as descriptive sentiment analyses. For temporal patterns, daily volume of breast screening tweets and daily average of composite sentiment scores were plotted in a line graph.

For geospatial patterns, tweets with available geographic information were used to generate point and hot spot maps based on composite sentiment scores. Hot spot analysis identifies spatial clusters with significantly high or low sentiment values, using the Getis-Ord Gi* statistics (170):

$$G_i^* = \frac{\sum_{j=1}^{n} w_{i,j}x_j - \bar{X}\sum_{j=1}^{n} w_{i,j}}{S\sqrt{\frac{\left[n\sum_{j=1}^{n} w_{i,j}^2 - (\sum_{j=1}^{n} w_{i,j})^2\right]}{n-1}}}, \text{ and}$$

$$\bar{X} = \frac{\sum_{j=1}^{n} x_j}{n}; S = \sqrt{\frac{\sum_{j=1}^{n} x_j^2}{n} - (\bar{X})^2}$$

where Gi* statistics is calculated on each location point i that has a feature (sentiment) value. The x_j is the sentiment value for features j, $w_{i,j}$ is the spatial weight between features i and j, and n is the total number of features. $\bar{X}$ is the average sentiment value from all features and S is the variance of sentiment values. Inverse square distance is used such that closer features are weighted more heavily than features that are further away. Let d_{ij} be the distance between features i and j, w_{ij} is equal to $M/(d_{ij}^2)$, where M is a constant. Conceptually, Gi* statistics compares the sum of feature values within a neighbouring region around location i against the expected sum of feature values derived from global average (numerator), and then standardized with the variance (denominator). The Gi* statistics returns a z-score for each location i. Significant hot spots contain highly positive z-score value and small p-value, indicating location i is surrounded by high sentiment value neighbours, while significant cold spots contain highly negative z-score and small p-value, indicating location i is surrounded low sentiment value neighbours.

For thematic patterns, a word-cloud was generated which consisted of the most frequent words amongst all negative tweets. A comprehensive list of common but non-informative words such as “the”, “it”, and “what” were omitted from word-cloud creation. The font size of each word in a word-cloud corresponded to the frequency of that word (i.e., the larger the word, the more frequently it appears). Example themes were extracted qualitatively as a demonstration. Sentiment map and word clouds for Canada were created and presented as supplementary figures: **Supplementary Figure 5-1**, **Supplementary Figure 5-2**, and **Supplementary Figure 5-3**.

5.2.5 Hypothesis-based sentiment analysis

To evaluate possible correlation between breast screening sentiment and actual breast screening uptake at an ecological level, a hypothesis-based sentiment analysis was conducted. While information on breast screening sentiment was provided by Twitter, information on breast screening uptake was obtained from a separate dataset collected by the (263) called Behavioral Risk Factor Surveillance System (BRFSS) survey. The BRFSS is one of the largest recurring national health surveys that collects data via phone interviews on U.S. residents regarding health-related risk behaviours, chronic health conditions, and use of preventive services. From the BRFSS 2014 survey (for calendar year 2014), interested individual-level variables were extracted and recoded as 1) mammogram received within the last two years (Mamm, 1 – yes and 0 – no), 2) CBE received within the last two years (CBE, 1 – yes and 0 – no), 3) highest education achieved (Edu, 1 – have at least some college education, 0 – do not have any college education), 4) general health (GenHlth, 1 – good, very good, or excellent, 0 – fair or poor), and 5) race (Race, 1 – non-Hispanic white only, 0 – all others). Women aged less than 40 years old, women with missing key variables (i.e., mammogram and CBE), and men were removed from the data. Explanatory and outcome variables were aggregated by states, where individual sentiment values were grouped as means by state (i.e., $\overline{\text{SentNeu}}$, $\overline{\text{SentPos}}$, $\overline{\text{SentNeg}}$, and $\overline{\text{SentCom}}$ as mean neutral, positive, negative, and composite sentiment score, respectively), and individual BRFSS variable values were aggregated as percentage of "1" for each state (i.e., %Mamm as percent women reported having a mammogram within two years, and likewise for %CBE, %Edu, %GenHlth, and %Race). We hypothesized that U.S. states with more positive sentiment score values towards breast cancer screening (as indicated by tweets) are more likely to have higher overall uptake of breast cancer screening. This hypothesis was examined qualitatively and

quantitatively. Qualitatively, the point maps of state-level breast screening sentiment and breast screening uptake patterns were compared. Quantitatively, since the values of the dependent variables (%Mamm and %CBE) fall between 0 and 1, beta regression model was used to statistically test the relationship between sentiment scores and mammogram/CBE uptake. States (including Hawaii, Vermont, and Montana) with less than 100 tweet count were excluded from the analysis.

In multivariate beta regressions, the outcome variable was either %Mamm or %CBE, and the explanatory variable of interest was one of the four mean VADER sentiment scores (i.e., $\overline{\text{SentNeu}}$, $\overline{\text{SentPos}}$, $\overline{\text{SentNeg}}$, or $\overline{\text{SentCom}}$) plus other covariates including %Edu, %GenHlth, and %Race to adjust for potential confounding. Beta regression assumes y_k (i.e., %Mamm or %CBE), where k=1,2,…, n_{state} (number of individual U.S. states), to be distributed in beta distribution and its probability density function is given as:

$$f(y;u,z) = \frac{\Gamma(z)}{\Gamma(uz)\Gamma\big((1-u)z\big)} y^{uz-1}(1-y)^{(1-u)z-1}$$

where Γ is a gamma function, and $0<y<1$, $0<u<1$, and $z>1$. Let p and q be the shape parameters of beta distribution. The u is the mean and z is the precision parameter, given as $u=p/(p+q)$ and $z=p+q$. The systematic component of beta regression is as follows:

$$g_1(E(y_k)) = g_1(u_k) = \beta_0 + \beta_1 x_{sentiment_k} + \beta_2 x_{education_k} + \beta_3 x_{general\_health_k} + \beta_4 x_{race_k}$$

where $E(y_k)=u_k$ is the expected y_k, or mean u_k, in each state. It is linearly linked to the explanatory variables via the logit link function, $g_1(u)=log(u/(1-u))$. The random component of beta regression states that the expected outcome $E(y_k)$ is distributed in beta distribution which is expressed as $y_k \sim Beta(u_k, z_k)$, where $g_2(z) = log(z)$. The estimation of β and z was done by maximum likelihood estimation.

5.3 Results

5.3.1 Descriptive analysis - Temporal patterns

There were 3,544,328 breast cancer-related tweets collected in the data collection period. The majority of Twitter users were from North America, specifically there were 61,524 tweets from the U.S. and related to breast screening, and 54,664 of these tweets contained geographic information allowing for map analysis. The baseline daily breast screening tweet volume fluctuated between 100 and 200, with an explosive volume started in the beginning of October (also Breast cancer awareness month) and then gradually declined back to baseline (**Figure 5-1**). For the remaining portions of this paper, "sentiment score" refers to "composite sentiment score" unless specified otherwise. There were 29,034 neutral ($-0.3 \leq$ sentiment score ≤ 0.3), 21,561 positive (sentiment score > 0.3), and 4,069 negative (sentiment score < -0.3) tweets. The daily average sentiment score was above the zero line during almost the entire period, indicating that the overall sentiment towards breast screening was neutral-to-positive (**Figure 5-1**).

5.3.2 Descriptive analysis - Geospatial patterns

Figure 5-2 depicts the location and sentiment polarity classification of each breast screening tweets. A larger volume of tweets was published in the eastern states, which coincided with states with higher population density (264).

Figure 5-3 shows hot spot analysis using individual composite sentiment scores, regions in red, orange, and yellow were statistically significant clusters of low sentiment value, whereas, regions in different shades of clue were significant clusters of high sentiment values.

Three quintile maps as average sentiment score, percent of recent mammogram, and percent of recent CBE by states are shown in **Figure 5-4**. Among these maps, there was a

general horizontal gradient of low-quintile regions to high-quintile regions from west to east across the country.

5.3.3 Descriptive analysis - Thematic patterns

The word-cloud constructed from negative breast screening tweets only is shown in **Figure 5-5**. Some of the key words were highlighted (bottom half of **Figure 5-5**) and could be grouped together thematically by a human inspector qualitatively. For example, “scared”, “pain”, “hurt”, and “discomfort” might be grouped together to suggest many people with negative sentiment about breast screening possibly concern about the physical and psychological discomfort with the procedure. On the other hand, “cost”, “insurance”, and “access” together might suggest some people viewed financial obstacles as a deterrence for breast screening.

5.3.4 Hypothesis-based sentiment analysis

Ecological correlations between each of the four average sentiment scores on breast screening and outcome variables (i.e., %Mamm and %CBE) were explored using multivariate beta regressions (**Table 5-1**). Significant positive correlation ($p<0.05$) was observed between $\overline{\text{SentNeu}}$ and recent CBE uptake, and significant negative correlations ($p<0.05$) were observed between $\overline{\text{SentNeu}}$ and recent mammogram uptake and recent CBE uptake. For example, 1% increase in average negative sentiment score was associated with 0.057 and 0.075 decrease in log odds of recent-mammogram-uptake being “yes” and recent-CBE-uptake being “yes”, respectively, adjusted for education, general health, and race.

5.4 Discussion

This study demonstrated how Twitter might serve as a potentially useful tool in fulfilling public health needs that require data on public perception. Twitter provides a rich source of

continuous, instantaneous, and uncensored public perception data, which may be utilized to monitor public sentiment towards health interventions. The descriptive sentiment analysis illustrated how Twitter depicts temporal, geospatial, and thematic patterns of sentiment. Temporally, the quantity and average sentiment typically fluctuated within a baseline range, which can help detect instances with abnormal level of tweet volume and/or sentiment score value. Point and hot spot maps visualized general geographical trends and specific clusters based on sentiment values, respectively. A vast number of negative sentiment in a location towards breast screening might indicate an underlying public misconception, unaddressed concerns, ineffective health promotion, or lack of accessible infrastructure. Thematically, qualitative interpretation of word-cloud revealed possibly important thematic elements that might lead to better understanding of the root causes of the observed sentiment in the whole country or specific regions.

In the hypothesis-based sentiment analysis, significant correlations were found between some of the mean sentiment scores (from Twitter) and actual mammogram and CBE uptake behaviour (from BRFSS 2014) at the state level. Average negative sentiment scores were negatively associated with mammogram and CBE uptakes, as expected. However, positive association was not observed between average composite and positive sentiment scores and breast screening uptakes. This might be due to several methodological and data-limitation challenges including data in Twitter and BRFSS did not overlap over the exact time period, subjects in these data sources did not represent the same individuals (i.e., Twitter users might not be representative to the target general population), relationship existed at the ecological (state) level could be different from individual level, uptake behaviours influenced by factors other than sentiment could be at play, certain states only had small number of tweets, and positive tweets

published by commercial or non-commercial organizations rather than individuals might not link to individuals' uptake patterns. Some of these Twitter data limitations were also mentioned by other studies such as (108), (265), and (266). Nonetheless, our finding suggested the existence of meaningful associations that negative sentiment tweets on breast screening might be particularly useful in identifying or predicting regions with lower breast screening uptake. Future studies are suggested to develop strategies to minimize background noise such as tweets published by organizations instead of individuals, and examine more fine-grained categorization of sentiment that also captures a person's feelings and moods such as anger, worry, disgust, fear, happiness, surprise, and sadness (267).

5.5 Conclusion

Based on the health belief model, one's perception about a health intervention could influence one's ultimate action in adopting it. Twitter sentiment data may fill an important role in providing health researchers and other stakeholders continuous and unfiltered data that is essential to gauge public perception on health interventions. The knowledge of such public perception might help predict subsequent public consumption. This study not only demonstrated the use of Twitter to visualize rich breast screening sentiment information, but also linked the sentiment derived from Twitter to actual breast screening uptake patterns from BRFSS 2014. This suggested that knowledge about public perception of health intervention may help predict future public consumption, which holds important values in public health policy development, community planning, and resource allocation. With better understanding and distillation of useful tweets from the noise, Twitter could potentially be used as a public health surveillance tool to monitor public perception. Spatial clusters with highly negative sentiment should be monitored closely over time and the reasons for their negative sentiment might be extracted

using thematic tools such as word-cloud. Specific programs or policies can be tailored to alleviating the negative sentiment, which might contribute to improving public's acceptance and consumption of a target health intervention.

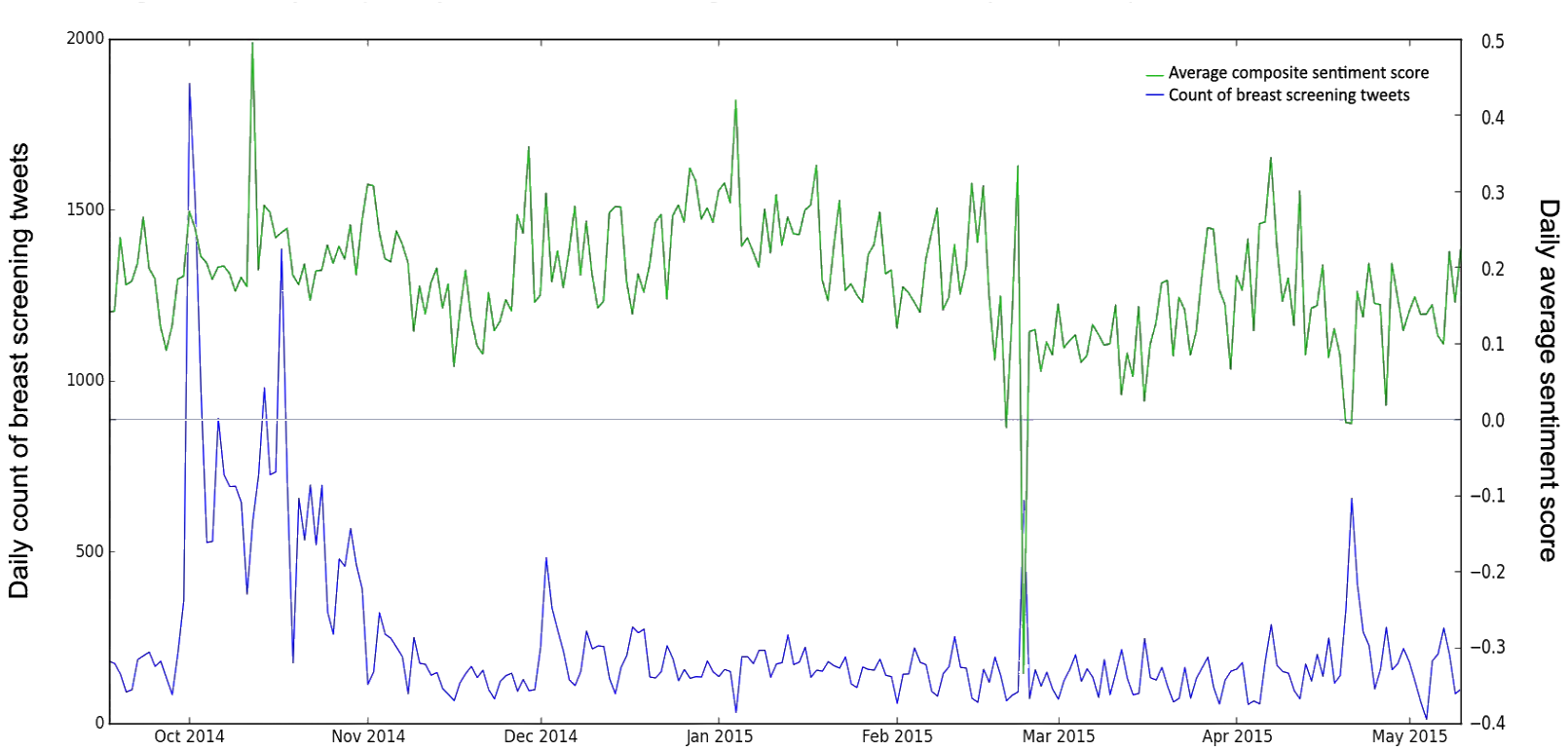


Figure 5-1: Daily average composite sentiment score and daily frequency of breast screening tweets in the U.S. (n_{tweet}=54,664).

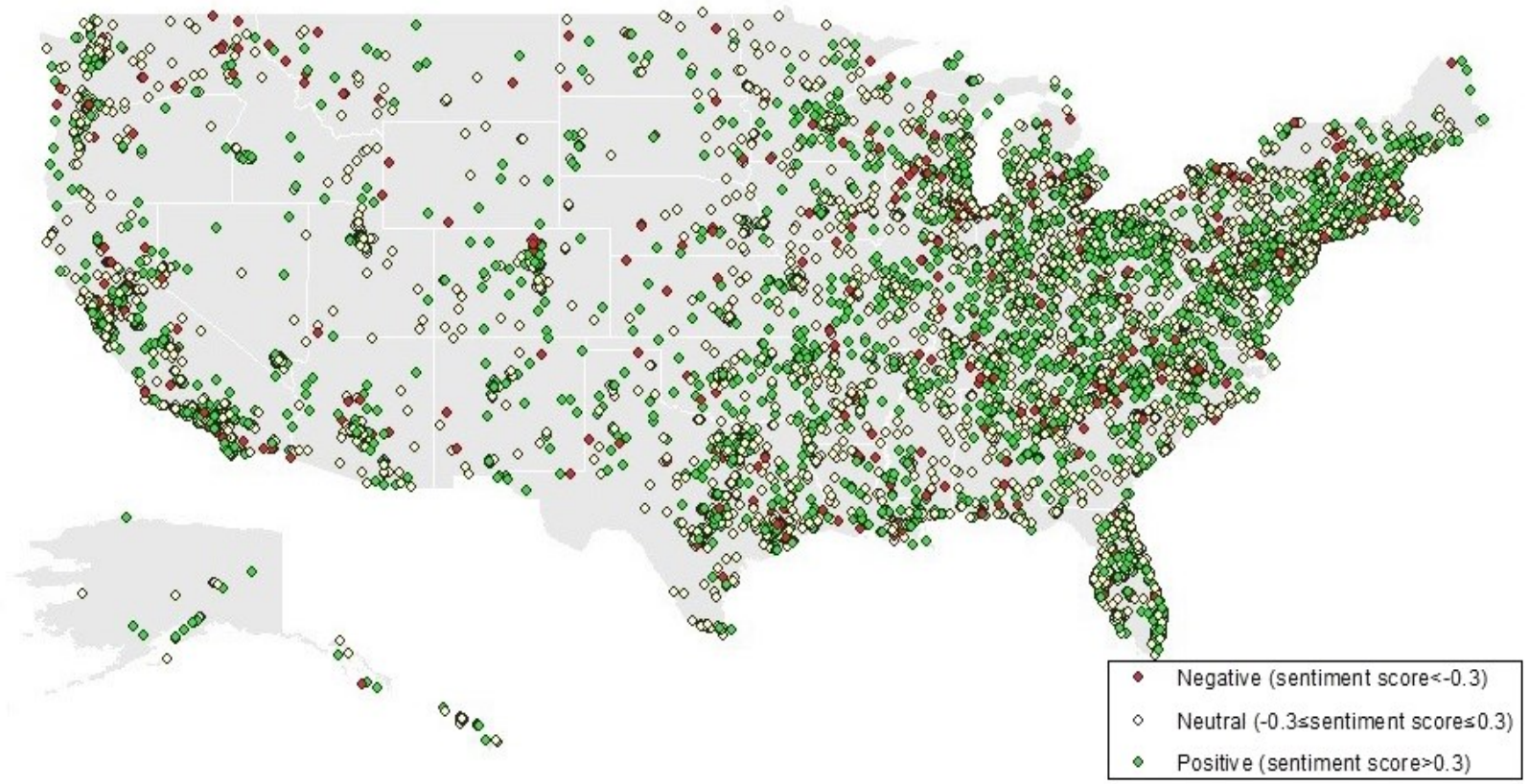


Figure 5-2: Sentiment of breast screening tweets in the U.S. (n_{tweet}=54,664).

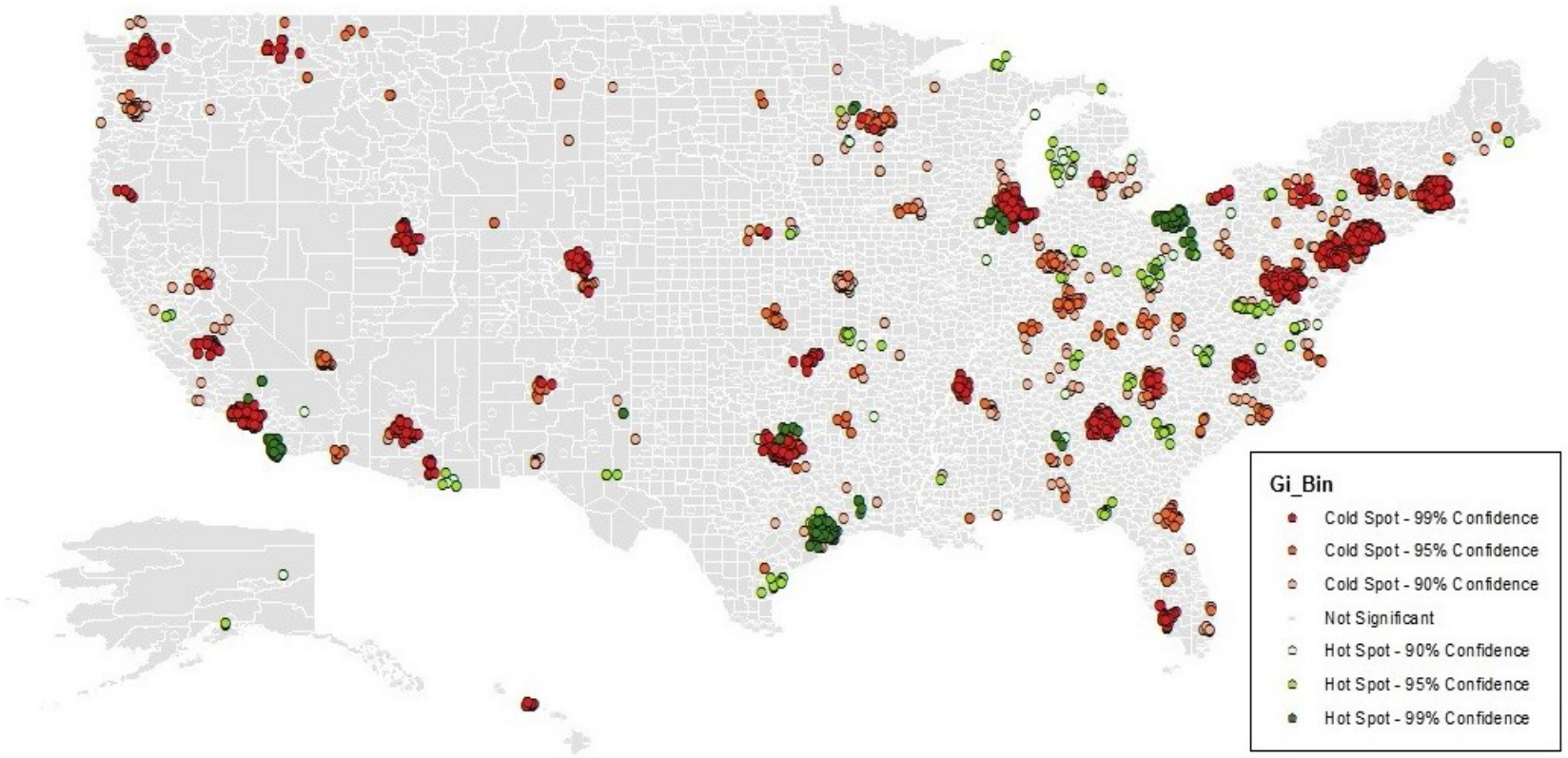


Figure 5-3: Hot spot map using composite sentiment score in the U.S. (n_{tweet}=54,664).

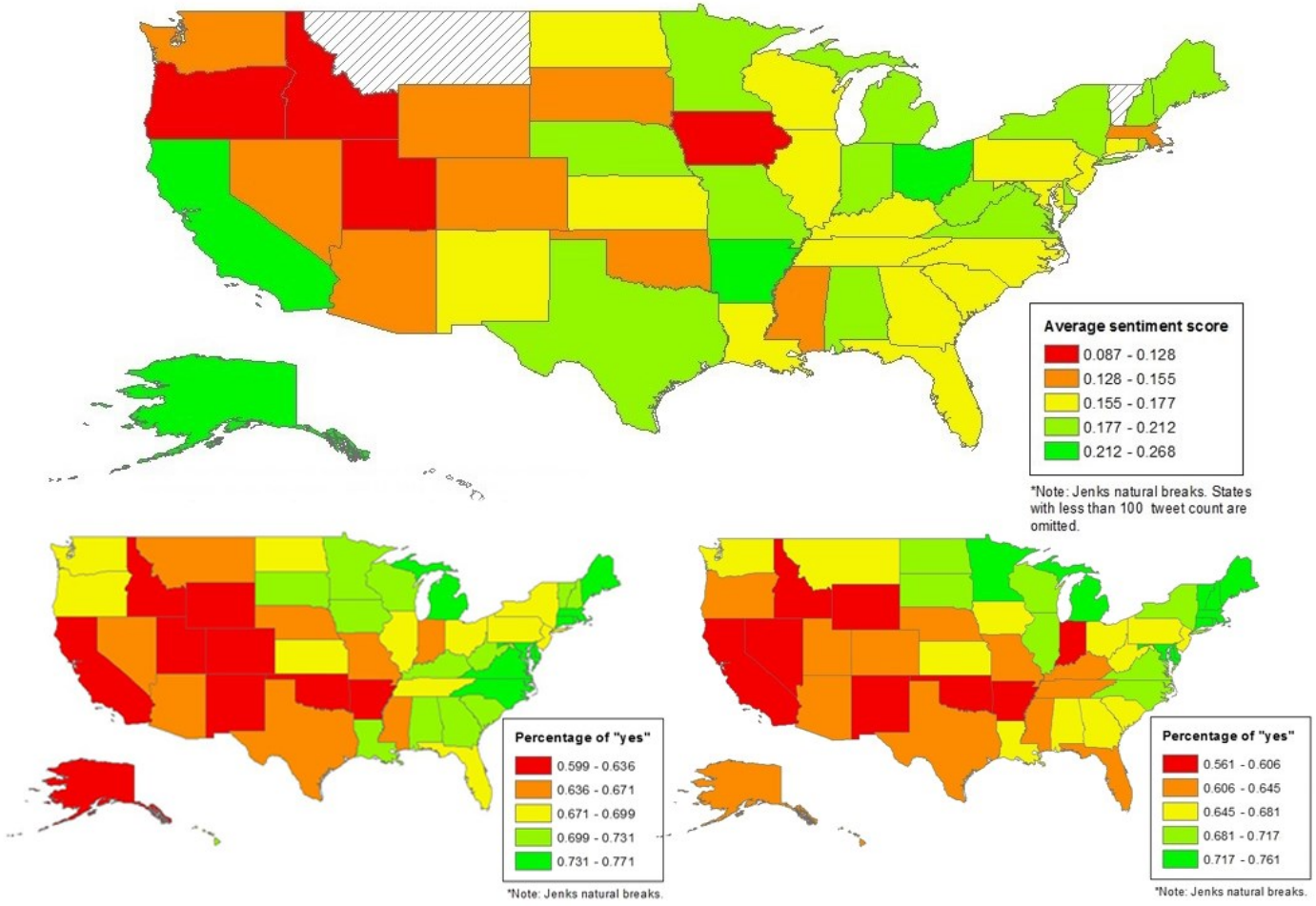


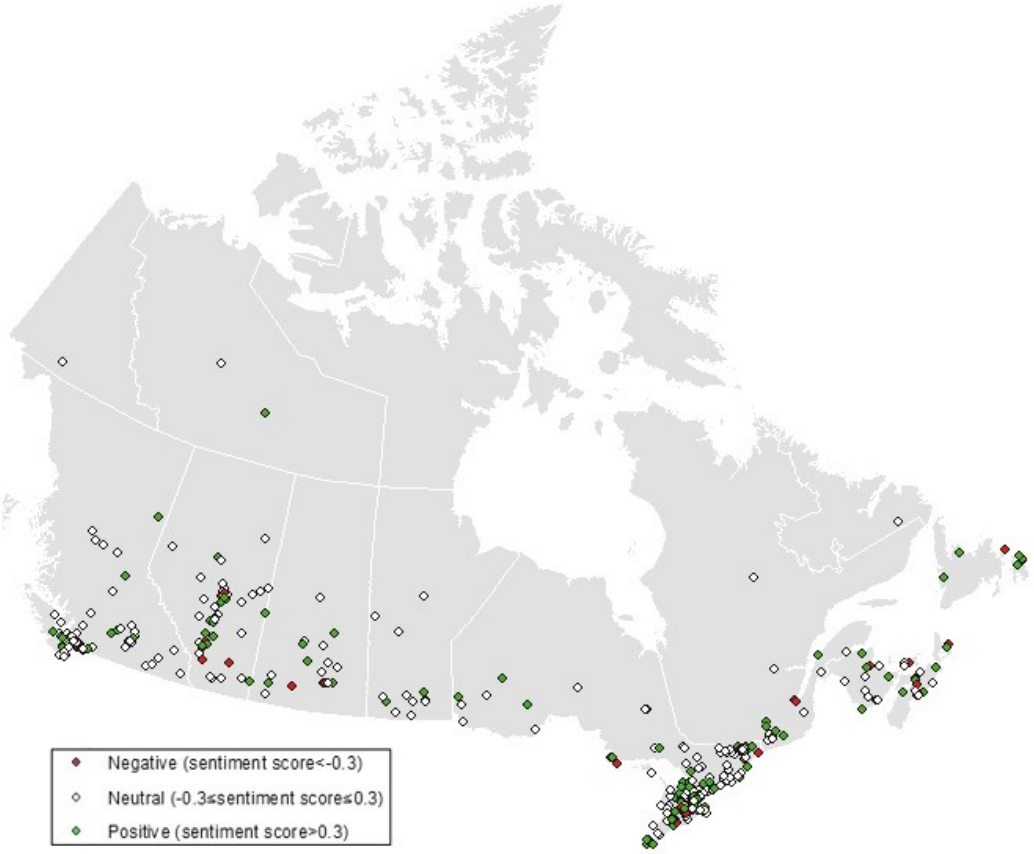
Figure 5-4: Quintile maps of average composite sentiment score of breast screening tweets (n_{tweet}=54,416, top), percent women aged ≥40 years with recent mammogram (n_{BRFSS}=217,503, bottom left), and percent women aged ≥40 years with recent CBE (n_{BRFSS}=217,503, bottom right).

Figure 5-5: Word-cloud using all negative breast cancer screening-related tweets in the U.S. (n_{tweet}=4,069).

Table 5-1: Multivariate beta regression examining average sentiment scores and outcome variables of recent mammogram and CBE uptakes by states (n_{state}=48).*

	$\overline{\text{SentCom}}$	$\overline{\text{SentNeu}}$	$\overline{\text{SentPos}}$	$\overline{\text{SentNeg}}$
%Mamm	-0.01 (-1.89 to 1.87)†	2.69 (-0.17 to 5.56)	-2.46 (-7.06 to 2.13)	-5.65 (-10.84 to -0.47)
%CBE	0.65 (-1.20 to 2.50)	3.27 (0.47 to 6.08)	-2.45 (-7.00 to 2.10)	-7.53 (-12.47 to -2.58)

*States with less than 100 tweets were removed. At the state level, $\overline{\text{SentCom}}$=average composite sentiment score, $\overline{\text{SentNeu}}$=average neutral sentiment score, $\overline{\text{SentPos}}$=average positive sentiment score, $\overline{\text{SentNeg}}$=average negative sentiment score, %Mamm=percent women (≥40 years) received mammogram within two years, and %CBE= percent women (≥40 years) who have received CBE within two years. †β-coefficient (95% C.I.) adjusted for education, general health, and race.



Supplementary Figure 5-1: Sentiment of breast screening tweets in Canada (n_{tweet}=2,821).

Only four jurisdictions had more than 100 breast cancer screening-related tweets over the study period (Alberta, British Columbia, Ontario, Quebec). Their average sentiment scores are: Alberta is 0.151 (n_{tweet}=308); British Columbia is 0.137 (n_{tweet}=714); Ontario is 0.160 (n_{tweet}=1,375); and Quebec is 0.144 (n_{tweet}=146).

Supplementary Figure 5-2: Word-cloud using all negative breast cancer screening-related tweets in Canada (n_{tweet}=222).

Supplementary Figure 5-3: Word-cloud using all positive breast cancer screening-related tweets in Canada (n_{tweet}=959).

Chapter 6 – Discussion

6.1 Key Findings and Discussions

6.1.1 Study I: Name- and location-based Aboriginal ethnicity classification

As an ethnically-diverse country, there is a general lack of ethnicity information in many public health-related data sources in Canada. The unavailability of ethnicity data may lead to researchers and policy-makers missing opportunities to identify health challenges, needs, and patterns amongst disadvantaged ethnic groups such as the Aboriginals. Our ethnicity study aimed at filling such Aboriginal ethnicity data gap existing in the majority of the databases by automating the prediction of one's Aboriginal status using simple and commonly-collected name and residential location information.

The study (presented in Chapter 4) has shown a number of important positive findings. Our Aboriginal ethnicity classifiers were deemed capable to classify the Aboriginal (all-inclusive) and First Nations groups reasonably accurately according to the performance metrics. The residential location feature appeared to be a necessary and critical feature to predict Aboriginal ethnicity, in conjunction to the name features. The location feature vastly improved the sensitivity values for Aboriginal and First Nations classifications, via the lowering of the volume of false negatives (Aboriginal individuals mistakenly identified as non-Aboriginals). Our study examined a simple location feature and a comprehensive set of 15 name features. For the classifications of Aboriginal and First Nations, the location feature tended to be more informative than name features (Chapter 4). Individuals living in the territories and certain regions in Alberta, Ontario, and Saskatchewan have higher likelihood to be classified as Aboriginal or First Nations. In terms of name features, many of the individual name features

were shown to be part of the top five most informative features (excluding the location feature), indicating that our derived list of 15 name feature set is comprehensive without high feature redundancy.

Besides Aboriginal (all-inclusive) and First Nations classifications, the classification performance for the rest of Aboriginal and sub-Aboriginal groups' classification were mixed. Even with the location feature, the sensitivity value for Métis classification remained poor. However, this is most likely due to the inherent difficulty of accurately and reliably inferring someone as Métis by his/her name since "Métis", by definition, refers to individuals with mixed Aboriginal and European heritages. The classification of Inuit and most of First Nations major Aboriginal language and tribal groupings were particularly challenging due to small sample size. The large improvement shown in many of these groups from using the full 1901 census, as opposed to 10% subsample, suggested that our classifiers for these groups could potentially achieve better performance once more samples were available.

The demonstrative application used our First Nations classifiers in the process of disease prevalence estimation within a number population/disease scenarios to cover a wide range of population disease settings. The use of high decision boundary (0.90), resulting in relatively higher PPV and lower sensitivity, tended to show the largest number of prevalence closely approximating (within the 95% C.I.) the underlying true prevalence. While the level of acceptable error between the estimated ethnicity-specific prevalence differs between different goals and objectives of the applications, applications intended to utilize automated Aboriginal ethnicity classifier may include a sensitivity analysis with various values of decision boundary in order to cover a range of estimated prevalence statistics.

6.1.2 Study II: Sentiment analysis on breast cancer screening tweets

Real-time online social media data with little to no restrictions on user input, such as Twitter, may provide important, unconventional, and previously-inaccessible information about the perception of the general public towards specific preventative public health interventions. In the first part of our sentiment study, we demonstrated a comprehensive data visualization as a potential public health surveillance tool by using breast cancer screening-related tweets to show temporal, geospatial, and thematic patterns pertaining to public perception towards the screening procedures in the U.S. These descriptive results seemed to be reliable based on a number of expected observations, such as having much higher volume of tweets during October (the breast cancer awareness month), average tweets about breast cancer screening falling between neutral and slightly positive, and higher tweet volume amongst the U.S. regions (the Eastern side of the country) with higher population density.

Hot spot maps identified geographic regions with significantly high or low sentiment towards breast cancer screening in the U.S. Thus, they may allow public health promoters, program developers, and policy-makers to better identify and focus on geographic regions that currently hold a widely negative perception towards a public health intervention. For example, significant cold spot regions may require change or improvement in specific health policies and education needed to address related physical, social, and systemic barriers. Our hot spot maps could potentially be generated in real-time (not shown in the study), which allows the estimation of the background noise and significant sentiment signals in various time frames, which is necessary for developing responsive and timely target interventions. Thematic analysis such as word-clouds based on the frequency of words in tweets in different sentiment polarity can enable

researchers, healthcare professionals, and decision-makers to obtain potential leads on the underlying reasons of why positive or negative sentiments exist for a particular region.

In the second part of this study, we have identified some significant association between breast cancer screening sentiment and actual breast cancer uptake from an external CDC data source (BRFSS). Based on the HBM, one's perception about a public health procedure can subsequently lead to corresponding action of the individuals (such as taking up or maintaining the procedure). Ideally, perception data preceding behavioural data will be more indicative of the possible causal pathway between perception and action/behaviour. However, due to data limitations, our Twitter data was derived from 2014-2015 while the BRFSS represented the 2014 calendar year. While significant ($p<0.05$) positive association was not found between the composite and positive sentiment scores and recent uptake (within two years) of mammogram and CBE, significant inverse relationships were identified between the negative sentiment score and uptake of mammogram and CBE. This may be due to the fact that a small portion of positive tweets was not published by individuals, but by commercial or non-commercial organizations advertising products related to or promoting the health benefits of breast cancer screening. Tweets from these organizations tended to contain positive, and not negative, sentiment. The existence of these positive tweets may dilute the underlying association between the positive sentiment scores and uptake behaviours. On the other hand, there may be an inherently stronger relationship between negative sentiment and associated uptake behaviours, than that with positive sentiment. This may suggest that negative sentiment may be a necessary and sufficient predictor for future uptake behaviour.

6.2 Public Health Significance

6.2.1 Study I: Name- and location-based Aboriginal ethnicity classification

Name/location-based ethnicity Aboriginal classifier can be used in a broad range of public health applications including subdividing population into groups of common origin, predicting Aboriginal compositions in residential areas, monitoring migration, studying healthy immigrant effect, detecting census undercount, measuring residential segregation, evaluating equal opportunity policies and political empowerment processes, and improving public and private services to Aboriginal populations (96, 97). The "predicted" Aboriginal ethnicity variable in previously-absent datasets will create the possibility to further study and reveal Aboriginal's health and social challenges in Canada with a fine spatial, temporal, and nominal granularity. Our name/location-based Aboriginal classifier will be a cost-efficient alternative when Aboriginal status information is absent, missing, or of low quality in existing databases (97).

In Canada, the ethnicity and Aboriginal ethnicity variable is collected only by a handful of surveys, including the Canadian Census, CHMS, CCHS, NPHS, APS, and NHS. Surveys that collect ethnicity/Aboriginal status information may suffer from at least one of the following limitations: lack of sufficient granularity (i.e. not enough Aboriginal subgroup options to select); low frequency of collection cycles, lack of standardized ethnicity-related survey questions and answering options between surveys, small sampling, and a degree of misclassification and non-participation (97). Clarke *et al.* (2008) stated that there is an absence of a "seminal Canadian paper that discusses the methodological issues related to the definition, conceptualization and operationalization of ethnicity and their implications, to guide future research and to enable comparability across studies" (90).

In contrast, personal names and corresponding residential location are recorded in most of the databases in Canada. Our study findings suggested that the prediction of the status of a number of Aboriginal and sub-Aboriginal groups (i.e., all-inclusive, First Nations, Algonquian, and Kootenay) was reasonably accurate. This automated classification method can potentially help unveil previously-unknown public health insights specific to Aboriginal individuals with little to no additional cost. This, in turn, may help health professionals and decision-makers to better develop, monitor, and implement culturally- and ethnically-competent policies and interventions. Aboriginal Canadians generally experience more social injustice (i.e., racism and income inequality) and disadvantaged health outcomes (i.e., disproportionately higher in many acute and chronic conditions) (100). By applying our Aboriginal ethnicity classifier on current health-related data sources, health and disease patterns can be stratified and compared between Aboriginal and non-Aboriginal groups. This may augment the effectiveness in monitoring Aboriginal-specific disease burdens and evaluating population health programs and policies that target in improving the health of Aboriginal Canadians. Culturally- and ethnically-appropriate programs and policies are more likely to be brainstormed, devised, and developed by having a clearer, finer, and more timely description of the population health/disease states of Aboriginal Canadians and their respective needs (268).

In addition, another public health significance of this study may go beyond Canada. Health disadvantages amongst the Aboriginal subpopulation and data challenges Aboriginal identification exist in other countries including Australia, New Zealand, and the U.S. (269). Our study is the first study that examined an automated ML approach to predict Aboriginal ethnicity which may serve as an analytic framework for researchers in other countries.

6.2.2 Study II: Sentiment analysis on breast cancer screening tweets

Health information derived from Twitter and other online social media differ from health information collected from traditional means such as surveys and administrative records. People tend to express their opinions on Twitter more freely and instantaneously (104, 105). Our study was considered one of the first studies in North America using ML-based methods to conduct a sentiment analysis on preventive cancer screening programs. We also explored the potential correlation between sentiment of breast cancer screening tests from Twitter and external CDC-published breast cancer screening uptake statistics from the BRFSS (270).

By using Twitter, public perception of breast cancer screening from large populations was summarized. A large volume of strongly negative sentiment towards breast cancer screening in a region could be an indication of the underlying public misconception, ineffective health promotion, or lack of accessible infrastructure in the region. The descriptive sentiment analysis illustrated the dynamics of sentiment towards breast cancer screening by time, space, and themes. Regions with prolonged and highly negative sentiment indicated areas should warrant further investigation as to what was mentioned in the negative discussions. Such visualization techniques demonstrating the sentiment patterns can potentially be incorporated into public health surveillance to help identify when, why, whom, and where may require mass interventions to promote better acceptance by the general public.

Based on the HBM, we believed that public perception about breast cancer screening tests may influence individual behavior in attending the tests. We postulated there should be positive associations between breast cancer screening sentiment and breast cancer screening uptake. While a similar study has shown positive relationship between Twitter-derived sentiments towards a novel H1N1 vaccine and CDC-estimated vaccination rates by region in the

U.S. (140), no study has been found to examine Twitter's use in a cancer prevention context. Knowing the sentiment towards preventive cancer screenings in a geographic region can help researchers, policy makers, and public health planners better identify specific "at-risk" communities, allocate needed resources, and develop appropriate programs and policies. Our study has identified a significant negative association between negative breast cancer screening sentiment and uptake of mammogram and CBE. This is particularly important and may allow prediction of future breast cancer screening uptake by assessing current perception towards these cancer screening procedures. This could be an important piece of information allowing healthcare professionals and decision-makers to better prepare appropriate and timely interventions to help remove such negative sentiments, as well as better prepare resources needed for the screening procedures themselves.

Our study suggested that public sentiment towards public health intervention may be an important risk factor that may be associated and/or potentially causally-linked with health outcomes. Public health researchers may devise, standardize, and promote measurement of sentiment-related information from channels other than online social media, including traditional survey and administrative databases. In addition, our proposed study also set the groundwork for future studies that are interested in utilizing perception mined from social media on different cancer types and cancer-related public health interventions.

6.3 Study Limitations

6.3.1 Study I: Name- and location-based Aboriginal ethnicity classification

Despite the use of the entire 1901 census in the analysis, large amount of class imbalance occurred with low frequency/representativeness of Aboriginal and its subgroups. The Aboriginal

(all-inclusive), FN, Métis, and Inuit consisted of only 1.89%, 1.22%, 0.47%, and 0.01% of the total sample, respectively. Examining FN's major language and tribal groupings were also challenging as breaking down into each sub-Aboriginal group made the representative sample even smaller. For example, each individual language and tribal FN group represented, on average, 0.15% of the total sample. Many of the sub-Aboriginal groups, such as Inuit, Athapaskan, Wakashan, Siouan, Tsimshian, Micmac, Iroquois, and Algonquin, that received a poor (<0.45) sensitivity at full census analysis can not be affirmed whether it was due to classifiers unable to learn the name/location patterns, no underlying name/location-ethnicity patterns existed, or not enough representative minority class sample for the classifiers to learn from.

A small degree (<3%) of typos and misspelling in name and ethnicity variable fields existed in the original 1901 census data. This would potentially lower the classifiers' ability to learn the underlying name-ethnicity rules. Furthermore, random noise due to subgroups of population who changed their names due to personal and cultural preferences (i.e., women's adoption of husband's last name upon marriage) will weaken the underlying association between name and Aboriginal ethnicity. Similarly, individuals changing residential location would weaken the association between location and Aboriginal ethnicity. Ethnicity is inherently a subjective concept that may change for an individual over time. These are the main data and methodologic challenges existed in our Aboriginal classifiers.

6.3.2 Study II: Sentiment analysis on breast cancer screening tweets

There is a degree of uncertainty of the population represented by the breast cancer screening-tweets collected for this study. Online social media users are a non-representative sample of the entire U.S. population. Social media users tended to be younger, living in urban

areas, and college-educated (271). In 2016, 24% of online adults and 21% of all adults in the U.S. used Twitter (272). There were 54% of all adult U.S. Internet users aged 30 years old and above used Twitter and 25% of U.S. adult women used Twitter. Since Twitter is not a direct random sample from our targeted population (U.S. women eligible for regular breast cancer screening), it may be prone to self-selection bias (273). Since the demographic profiles of total Twitter users and those users who post breast cancer-related tweets are unknown, the extent of impact of such self-selection bias can not be evaluated. The potential self-selection bias may affect our study's generalizability to the target population. Furthermore, our collected tweet data may contain a small amount of tweets not published by real individuals, but by organizations or bots. However, the quantity of them appeared to be negligible (<1-3%) when manually examining a random subset of collected tweets (N=250).

The association analysis in the hypothesis-driven analysis section has shown some expected association between breast cancer screening sentiment and screening uptake behaviour. Ideally, the sentiment and uptake information should be extracted from the same individuals in a reasonable chronological order by collecting sentiment information before uptake information. This will help establish a clearer temporality element based on the Bradford Hill's causal criteria (274). Our study was also constrained by the availability of data sources as no single data source was able to provide both sentiment and screening uptake information simultaneously at the individual level. Thus, sentiment was derived from Twitter and uptake behaviour was derived from BRFSS, and then linked/aggregated at the state level. At the time of the study, the beginning of tweet collection started in September 2014 and ended in April 2015, and the most up-to-date BRFSS available in 2015 was derived from calendar year 2014. Since the individuals consisted of these two datasets are different and can not be cross-linked, establishing a clear

causal relationship between individual perception and individual screening uptake behaviour at individual level is infeasible.

6.4 Recommendations on Future Studies

6.4.1 Study I: Name- and location-based Aboriginal ethnicity classification

Our study is one of the first and most comprehensive name/location-ethnicity classification studies for predicting Aboriginal and sub-Aboriginal statuses in Canada. Future studies are recommended to target and expand on the aforementioned limitations pertaining to our study. To address the low number of trainable minority class sample, multiple years of Canadian censuses and/or other data sources (i.e., CCHS and APS) providing name, location, and Aboriginal ethnicity can be combined. However, attention to corresponding data processing/standardization may be needed when combining multiple years or multiple data sources since the definitions, framing questions, and available response type and categorization are likely to be different. In addition, samples in data sources collected within a small time frame may contain duplicate cases, appropriate methods such as identifying/eliminating duplicate or suppressing their data representation by lowering their weights may be needed. On the other hand, the classification performance derived from the use with more recent data sources may give a more generalizable inference to how good the classifiers for modern time. Furthermore, additional features paired with a feature reduction method can be used to explore new name and location feature sets while minimizing the likelihood of overfitting. An Aboriginal language linguist/expert can be included to provide deeper insights into the naming systems/practices of Aboriginals in order to possibly generate additional name features. Furthermore, advanced ML algorithms such as neural network and ensemble algorithms (i.e., random forest) may be explored.

Further validation studies to not only validate the accuracy of classifying Aboriginal status but also examine misclassification's impact on population health and disease statistics should be conducted. For example, the CCHS contains Aboriginal identifier and chronic disease conditions. Predicted Aboriginal status (i.e., First Nations, Métis, Inuit, and non-Aboriginal) may be computed using our name/location-based ethnicity classifiers. Aboriginal-specific disease statistics can be computed with the predicted Aboriginal status, and that can be compared directly with the true Aboriginal-specific disease statistics using the CCHS. A set of different health and disease conditions should be examined to grasp what disease settings the automated Aboriginal classifiers perform acceptably. Sensitivity analysis may be added to examine a wide range of decision boundary values. Caution must be taken applying this approach since the CCHS has its own data limitations such as small sampling frame and low response rates in certain regions.

6.4.2 Study II: Sentiment analysis on breast cancer screening tweets

Our study is one of the first studies demonstrating the potential to use Twitter as a public health surveillance tool for cancer screening procedures. Future studies are recommended to address some of the aforementioned study limitations inherent to Twitter data and its data processing procedures. To retain relevant tweets posted only by individuals, a user filtering step should be implemented to automatically detect and filter out spam tweets and promotional tweets by commercial and non-commercial organizations (275). Methods attempting to reduce the self-selection and sampling biases in Twitter data should also be explored. An illustration of this is shown by Culotta (2014) who inferred demographic attributes (such as sex and race) of Twitter users based on their communication patterns, and then visualized the data based on inferred demographics, using standard survey weighting (273). However, automated demographic

labeling is imperfect and may lead to dismissal of larger number of tweets that can not be automatically annotated. Furthermore, reweighting can also be explored on duplicated tweets or retweets since these tweets should not be weighted equally to original tweets in certain applications.

While Hutto and Gilbert's VADER (2014) sentiment classifier appeared to achieve a reasonably-high accuracy when predicting the sentiment on our breast cancer screening tweets, it certainly has room to improve. The version used in our study was 0.5, but the newest version 2.5 has been published in 2016 (276). Other studies may also develop their own sentiment classifiers with lexicons specifically attuned to public health domains. A number of researches has explored the creation of health-related sentiment lexicon which may be examined and/or built upon for future studies (277, 278). More sophisticated methods have attempted to differentially detect the holder of a sentiment (i.e., person) and the target (i.e., public health intervention) (279). Multi-dimensional sentiment analysis can be carried out to explore not just the simplistic representation of emotion as merely neutral, positive and negative, but in multiple dimensions including aspects of intent, risk, truthfulness/deception, and a wider range of emotions such as basic (i.e., anger, disgust, fear, happiness, and sadness), generic (i.e., abandon and affect), relational (i.e., abhor and love), caused (i.e., afraid and amused), causative (i.e., affront and offend), goal (i.e., covet and curious), and complex (i.e., ashamed and assured) (280). Future studies may also explore other public health interventions and/or disease conditions, such as examining the relationship between breast cancer screening sentiment and breast cancer incidence, or relationship between breast cancer screening sentiment and amount of breast cancer screening promotional/educational activities by regions.

To explore and consolidate the possible causal relationship between public sentiment and subsequent uptake behaviour of public health intervention, future studies may validate this possible causal pathway with a clearer temporal timeline between the expressed sentiment and subsequent uptake behaviour. Appropriate data sources or linkage may be needed to conduct such study at an individual level. Cause-and-effect has been shown within the Twitter sphere for commercial products (281). Twitter serves as a quick microblogging word-of-mouth platform that speeds up the dissemination of post-purchase product evaluations from consumers, thus allowing early product adoption and continual product promotion. Scarce research has been done to assess if sentiment and adoption of public health products (i.e., education, promotion, intervention, policy, and campaign) spread and uptake in similar fashion. Such knowledge may hold important value for public health by not only expanding the our understanding regarding the flow of sentiment and uptake information pertaining to public health products, but also advancing testable and practical strategies to better promote the adoption of healthy behaviours via online channels with large user base.

6.5 Final Remarks

This doctoral thesis aimed to demonstrate the relevance of big data and machine learning methods on public health settings by applying ML methods on traditional database (*Name- and Location-based Aboriginal Ethnicity Classification*) and utilizing unconventional real-time streaming online social media data (*Sentiment Analysis of Breast Cancer Screening in the United States Using Twitter*). These independent studies illustrated the large, untapped potential and corresponding data and methodological challenges of utilizing big data and ML methods to tackle contemporary public health issues in North America.

The ethnicity classification study illustrated that with simple and commonly-collected variables, Aboriginal ethnicity can be accurately predicted and may be able to fill the Aboriginal ethnicity gap for vast public health applications. This automated ethnicity prediction method can be leveraged by many stakeholders once further validated across a wider setting of Canadian data systems. The predicted ethnicity can be applied to better direct public health research in identifying Aboriginal-specific disease statistics which are essential for developing ethnically-appropriate health policies and interventions.

The sentiment study utilized fast, massive, and streaming data derived from Twitter and has shown a potential to visualize various descriptive patterns. It has identified meaningful associations between breast screening sentiment and uptake at the state level. This study can be used to supplement existing public health surveillance in which uncensored perception/sentiment-based information is not widely collected. Individuals' perception is a major foundation for the failure or success of many public health programs, yet public health researchers and professionals are often oblivious to such subjective information due to data unavailability. Thus, being able to accurately and reliably monitor mass perception towards public health interventions in a timely and economic fashion, as shown in our study, can be vital for policy, research, and program development.

New sources and types of big data and ML techniques will continue to emerge at a rapid pace. The field of public health and epidemiology should incorporate education and training to prepare their scientists who are interested in exploring and analyzing health-related big data. This thesis assisted in this regard by demonstrating many steps in big data analytics including data collection (i.e., via web indexing and accessing public API), data cleaning and processing, using existing resources (i.e., VADER sentiment classifier and open-source Python libraries), feature

generation, data visualization, cross-database validation, and uses of regression-based and ML-based techniques. With more research done using big data analytics in healthcare and public health, its potential and utility may become more fully realized. While big data and ML techniques have their own limitations, they provide a large additional repertoire of powerful quantitative tools for epidemiologists to handle data in more unconventional ways and to create more research and application opportunities.

References

1. Ingebrigtsen N. The differences between data, information and knowledge 2016 [cited 2016 12-12-2016]. Available from: http://www.infogineering.net/data-information-knowledge.htm.
2. Marr B. A brief history of big data everyone should read: LinkedIn; 2015 [cited 2016 12-12-2016]. Available from: https://www.linkedin.com/pulse/brief-history-big-data-everyone-should-read-bernard-marr.
3. Hilbert M. Course - digital technology and social change: University of California; 2015 [cited 2017 15-02-2017]. Available from: https://canvas.instructure.com/courses/949415.
4. OxfordDictionaries. Definition of information explosion in English: Oxford Unversity Press; 2017 [cited 2017 15-02-2017]. Available from: https://en.oxforddictionaries.com/definition/information_explosion.
5. Pfeiffer D, Stevens K. Spatial and temporal epidemiological analysis in the big data era. Preventive Veterinary Medicine. 2015;122:213-20.
6. Witten I, Frank E, Hall M. Data mining: practical machine learning tools and techniques. 3 ed. Burlington: Elsevier; 2011.
7. Marr B. The best 10 big data quotes of all times: LinkedIn; 2015 [cited 2016 12-12-2016].
8. International Business Machines Corporation. What is big data? 2013 [cited 2016 13-12-2016]. Available from: https://www-01.ibm.com/software/data/bigdata/what-is-big-data.html.
9. Villars R, Olofson C, Eastwood M. Big data: what is it and why you should care. IDC Analyze the Future, 2011.
10. Sanou B. ICT facts and figures: the world in 2015. Geneva, Switzerland: ICT Data and Statistics Division Telecommunication Development Bureau, 2015.
11. Vermesan O, Friess P, Guillemin P, Sundmaeker H, Eisenhauer M, Moessner K, et al. Internet of things strategic research and innovation agenda: River Publishers; 2013. 348 p.
12. Mell P, Grance T. The NIST definition of cloud computing: recommendations of the National Institute of Standards and Technology. U.S. Department of Commerce, 2011 Contract No.: Special Publication 800-145.
13. Reddi S. 4 ways big data will transform business: CSC Big Data and Analytics group; 2013 [cited 2016 13-12-2016]. Available from: http://www.csc.com/big_data/publications/89362/96477-4_ways_big_data_will_transform_business.
14. Kudyba S, Kwatinetz M. Introduction to the big data era. 2012.
15. Press G. A very short history of big data: Forbes; 2013 [cited 2017 06-01-2017]. Available from: http://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/#5360e6c555da.
16. Import.io. Everything you ever wanted or needed to know about big data 2015 [cited 2016 30-12-2016]. Available from: https://www.import.io/post/101-guide-to-big-data/.
17. Hurwitz J, Nugent A, Halper F, Kaufman M. Big data For dummies2013.
18. Grimes S. Is "unstructured" data merely unmodeled? : Intelligent Enterprise; 2015 [cited 2017 07-01-2017]. Available from: http://www.informationweek.com/software/information-management/structure-models-and-meaning/d/d-id/1030187?
19. Jin X, Wah B, Cheng X, Wang B. Significance and challenges of big data research. Big Data Research. 2015;2:59-64.
20. Gaffney B, Dobbs D, Health L. What is big data? HIMSS Clinical & Business Intelligence Data and Analytics Task Force, 2014.
21. Kuo M, Sahama T, Kushniruk A, Borycki E, Grunwell D. Health big data analytics: current perspectives, challenges and potential solutions. International Journal of Big Data Intelligence. 2014;1(1/2):114-26.

22. Zaïane O. Rich data: risks, issues, controversies and hype. Keynote at the 10th International Conference on Advanced Data Mining and Applications. 2014.
23. Marr B. Why only one of the 5 Vs of big data really matters 2015 [cited 2016 13-12-2016]. Available from: http://www.ibmbigdatahub.com/blog/why-only-one-5-vs-big-data-really-matters.
24. Gange S, Golub E. From smallpox to big data: the next 100 years of epidemiologic methods. American Journal of Epidemiology. 2015;183(5):423-6.
25. Mooney S, Westreich D, El-Sayed A. Epidemiology in the era of big data. Epidemiology. 2015;26(3):390-4.
26. Sun J, Reddy C. Big data analytics for healthcare. IBM; 2013.
27. Crown W. Potential application of machine learning in health outcomes research and some statistical cautions. Value in Health. 2015;18:137-40.
28. Goldstein B, Navar A, Carter R. Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges. European Heart Journal. 2016.
29. Samuel A. Some studies in machine learning using the game of checkers. International Business Machines Corporation, 1959.
30. Yoo I, Alafaireet P, Marinov M, Pena-Hernandez K, Gopidi R, Chang J, et al. Data mining in healthcare and biomedicine: a survey of the literature. Journal of Medical Systems. 2012;36(4):2431-48.
31. Donalek C. Supervised and unsupervised learning 2011 [cited 2017 14-01-2017]. 69]. Available from: http://www.astro.caltech.edu/~george/aybi199/Donalek_Classif.pdf.
32. Talwar A, Kumar Y. Machine learning: an artificial intelligence methodology. International Journal Of Engineering And Computer Science. 2013;2(12):3400-4.
33. Chen E. Choosing a machine learning classifier 2011 [cited 2017 13-01-2017]. Available from: http://blog.echen.me/2011/04/27/choosing-a-machine-learning-classifier/.
34. Multiple. When to choose which machine learning classifier? : StackOverflow; 2010 [cited 2017 13-01-2017]. Available from: http://stackoverflow.com/questions/2595176/when-to-choose-which-machine-learning-classifier.
35. Markham K. Comparing supervised learning algorithms: Data School; 2015 [cited 2017 06-02-2017]. Available from: http://www.dataschool.io/comparing-supervised-learning-algorithms/.
36. Magdon-Ismail M. Lecture 11: overfitting. Learning from data 2012.
37. Ye F, Zhao A. What you see may not be what you get – a brief introduction to overfitting 2010 [cited 2016 02-03-2016]. Available from: http://www.vicc.org/biostatistics/2010workshop/Overfitting_0416.pdf.
38. Caruana R. Bias/variance tradeoff. Empirical methods in machine learning & data mining 2005.
39. Roweis S, Saul L. Nonlinear dimensionality reduction by locally linear embedding. Science 2000;290(5500):2323–6.
40. Schneider J. Cross Validation. School of Computer Science, Carnegie Mellon University; 1997.
41. Polikar R. Ensemble based systems in decision making. IEEE Circuits and Systems Magazine. 2006;6(3):21–45.
42. Samet H. Foundations of Multidimensional and Metric Data Structures: Morgan Kaufmann; 2006.
43. Schneider J. Cross validation 1997 [cited 2017 22-03-2017]. Available from: https://www.cs.cmu.edu/~schneide/tut5/node42.html.
44. Shmueli G. To explain or to predict? Stat Sci. 2010;25(3):289-310.
45. Fortmann-Roe S. Understanding the bias-variance tradeoff 2012 [cited 2017 22-03-2017]. Available from: http://scott.fortmann-roe.com/docs/BiasVariance.html.
46. School of Public Health. What is epidemiology? : University of Alabama Birmingham; 2012 [cited 2017 08-01-2017]. Available from: http://www.soph.uab.edu/epi/academics/studenthandbook/what.

47. Bristish Medical Journal. Chapter 1. What is epidemiology? 2015 [cited 2017 07-01-2017]. Available from: http://www.bmj.com/about-bmj/resources-readers/publications/epidemiology-uninitiated/1-what-epidemiology.
48. Krieger N. Proximal, distal, and the politics of causation: what's level got to do with it? American Journal of Public Health. 2008;98(2):221-30.
49. New Health Advisor. Types of epidemiological studies 2014 [cited 2017 08-01-2017]. Available from: http://www.newhealthadvisor.com/Types-of-Epidemiological-Studies.html.
50. Management Study Guide. Secondary data: MSG; 2012 [cited 2017 14-01-2017]. Available from: http://www.managementstudyguide.com/secondary_data.htm.
51. Centers for Disease Control and Prevention. Ten great public health achievements - United States, 2001-2010 2011 [cited 2017 08-01-2017]. Available from: https://www.cdc.gov/mmwr/preview/mmwrhtml/mm6019a5.htm.
52. Higgins J, Strange K, Scarr J, Pennock M, Barr V, Yew A, et al. "It's a feel. That's what a lot of our evidence would consist of ": public health practitioners' perspectives on evidence. Evaluation and the Health Professions. 2011;34(3):278-96.
53. LaPelle N, Luckmann R, Simpson E, Martin E. Identifying strategies to improve access to credible and relevant information for public health professionals: a qualitative study. BioMed Central Public Health. 2006;6(1):89-101.
54. Keough K. The Third Amyot Lecture. How science informs the decisions of government. Canadian Journal of Public Health. 2002;93(2):104-8.
55. Ola O, Sedig K. The challenge of big data in public health: an opportunity for visual analytics. Online Journal of Public Health Informatics. 2014;5(3):e223.
56. Revere D, Turner A, Madhavan A, Rambo N, Bugni P, Kimball A, et al. Understanding the information needs of public health practitioners: a literature review to inform design of an interactive digital knowledge management system. Journal of Biomedical Informatics. 2007;40(4):410-21.
57. Rambo N. Information resources for public health practice. Journal of Urban Health. 1998;75(4):807-25.
58. Cohen B, Franklin S, West J. Perspectives on the Massachusetts Community Health Information Profile (MassCHIP): developing an online data query system to target a variety of user needs and capabilities. Journal of Public Health Management and Practice. 2006;12(2):155-60.
59. Reeder B, Revere D, Hills R, Baseman J, Lober W. Public health practice within a health information exchange: information needs and barriers to disease surveillance. Online Journal of Public Health Informatics. 2012;4(3):e4.
60. Kiefer L, Frank J, DiRuggiero E, Dobbins M, Manuel D, Gully P, et al. Fostering evidence-based decision-making in Canada: examining the need for a Canadian population and public health evidence centre and research network. Canadian Journal of Public Health. 2005;96(3):11-40.
61. Canadian Public Health Association. 12 great achievements 2012 [cited 2017 08-01-2017]. Available from: http://www.cpha.ca/en/programs/history/achievements.aspx.
62. Neumann P, Jacobson P, Palmer J. Measuring the value of public health systems: the disconnect between health economists and public health practitioners. American Journal of Public Health. 2008;98(12):2173-80.
63. Canada Health Infoway. Canadian physicians can improve patient care with advanced EMR use 2016 [cited 2017 10-01-2017]. Available from: https://www.infoway-inforoute.ca/en/what-we-do/news-events/newsroom/2016-news-releases/6752-canadian-physicians-can-improve-patient-care-with-advanced-emr-use.
64. Canadian Institute for Health Information. Better information for improved health: a vision for health system use of data in Canada. Ottawa, Ontario: Canadian Institute for Health Information, 2013.
65. Roy S. Big data analytics in healthcare. Octocube Consulting, 2014.

66. Canada Health Infoway. Big data analytics in health: white paper (full report). 2013.
67. Khoury M. Planning for the future of epidemiology in the era of big data and precision medicine. American Journal of Epidemiology. 2015;182(12):977–9.
68. Simon P. Too big to ignore: the business case for big data: Wiley and SAS Business Series; 2013.
69. Shortliffe E. Strategic action in health information technology: why the obvious has taken so long. Health Affairs (Millwood). 2005;24(5):1222-33.
70. Greenes R, Barnett G, Klein S, Robbins A, Prior R. Recording, retrieval and review of medical data by physician-computer interaction. The New England Journal of Medicine. 1970;282(6):307-15.
71. Consultants to Government and Industries. Data driven healthcare: the Canadian experience: CGI Health; 2015 [cited 2017 16-01-2017]. Available from: http://www.e-healthconference.com/pastpresentations/2015/201571435822/1960eHealth_2014_Data_Driven_Decisions_in_Canadian_Healthcare_v1.pdf.
72. Solomon H. Canadian health care analytics 'still in teen years,' conference told: IT World Canada; 2013 [cited 2017 16-01-2017]. Available from: http://www.itworldcanada.com/article/canadian-health-care-analytics-still-in-teen-years-conference-told/87149.
73. Closson T. The goldilocks principle and Canadian health care system governance 2014 [cited 2017 10-01-2017]. Available from: http://healthydebate.ca/opinions/the-goldilocks-principle-and-canadian-healthcare-system-governance.
74. Susser M. Does risk factor epidemiology put epidemiology at risk? Peering into the future. Journal of Epidemiology and Community Health. 1998;52:608-11.
75. Kunkle S, Christie G, Yach D, El-Sayed A. The importance of computer science for public health training: an opportunity and call to action. Journal of Medical Internet Research Public Health and Surveillance. 2016;2(1):e10.
76. Hutto C, Gilbert E. VADER: a parsimonious rule-based model for sentiment analysis of social media text. Eighth International AAAI Conference on Weblogs and Social Media. 2014.
77. Ularu E, Puican F, Apostu A, Velicanu M. Perspectives on Big Data and Big Data Analytics. Database Systems Journal. 2012;3(4):3-14.
78. Pentland A, Reid T, Heibeck T. Revolutionizing medicine and public health. Big Data and Health Working Group, 2013.
79. Breiman L. Statistical modeling: the two cultures. Statistical Science. 2001;16(3):199-231.
80. Afkhami R, Acik-Toprak N. Ethnicity: introductory user guide. The Economic and Social Data Service (ESDS) Government, 2012.
81. Anthony D, Baggott R, Tanner J, Jones K, Evans H, Perkins G, et al. Health, lifestyle, belief and knowledge differences between two ethnic groups with specific reference to tobacco, diet and physical activity. Journal of Advanced Nursing. 2011;68(11):2496-503.
82. Gottlieb N, Green L. Ethnicity and lifestyle health risk: some possible mechanisms. American Journal of Health Promotion. 1987;2(1):37-51.
83. Sundquist J, Johansson S. The influence of socioeconomic status, ethnicity and lifestyle on body mass index in a longitudinal study. International Epidemiological Association. 1998;27:57-63.
84. Chávez A, Guido-DiBrito F. Racial and ethnic identity and development. An update on adult development theory: Jossey-Bass; 1999. p. 39-47.
85. Tsai J, Chentsova-Button Y, Wong Y. Why and how researchers should study ethnic identity, acculturation, and cultural orientation. Asian American psychology: the science of lives in context: American Psychological Association; 2002. p. 41-65.
86. Comstock R, Castillo E, Lindsay S. Four year review of the use of race and ethnicity in epidemiologic and public health research. American Journal of Epidemiology. 2004;159(6):611-9.
87. Mays V, Ponce N, Washington D, Cochran S. Classification of race and ethnicity: implications for public health. Annual Review of Public Health. 2003;24(83-110).

88. Nestel S. Colour coded health care: the impact of race and racism on Canadians' health. Wellesley Institute, 2012.
89. Beiser M. The health of immigrants and refugees in Canada. Canadian Journal of Public Health. 2005;96(2):S29-S44.
90. Clarke C, Colantonio A, Rhodes A. Ethnicity and mental health: conceptualization, definition and operationalization of ethnicity from a Canadian context. Chronic Diseases in Canada. 2008;28(4):128-47.
91. Hyman I. Racism as a determinant of immigrant health. Ottawa, ON: Public Health Agency of Canada, 2009.
92. Lofters A, Shakardass K, Kirst M, Quinonez C. Sociodemographic data collection in healthcare settings: an examination of public opinions. Medical Care. 2011;49(2):193-9.
93. Parsons R. Institutionalized racism and classism: a meta analysis of Canadian and American studies of breast cancer care. Windsor, ON: School of Social Work, University of Windsor, 2005.
94. Varcoe C, Browne A, Wong S, Smye V. Harms and benefits: collecting ethnicity data in a clinical context. Social Science and Medicine. 2009;68:1659-66.
95. Randall V. Eliminating racial discrimination in health care. Totowa, NJ: Humana Press; 2007.
96. Mateos P. A review of name-based ethnicity classification methods and their potential in population studies. Population, Space and Place. 2007;13:243-63.
97. Mateos P. Name, ethnicity and populations: tracing identity in space: Springer; 2014.
98. Loppie S, Reading C, deLeeuw S. Social determinant of health: Aboriginal experiences with racism and its impacts. National Collaborating Centre for Aboriginal Health, 2014.
99. Canadian Human Rights Commission. Report on equality rights of Aboriginal people. CHRC, 2011.
100. Reading C, Wien F. Health inequalities and social determinants of Aboriginal Peoples' health. National Collaborative Centre for Aboriginal Health, 2009.
101. NCCAH. An overview of Aboriginal health in Canada. National Collaborative Centre for Aboriginal Health, 2013.
102. Public Health Agency of Canada. Diabetes in Canada: facts and figures from a public health perspective. Ottawa: Public Health Agency of Canada, 2011.
103. Smith K. 44 Twitter Statistics for 2016 2016 [cited 2017 18-01-2017]. Available from: https://www.brandwatch.com/blog/44-twitter-stats-2016/.
104. Zhao D, Rosson M. How and why people Twitter: the role that micro-blogging plays in informal communication at work. 2009:243-52.
105. Dredze M, Paul M, Lamb A, Broniatowski D. Social media: a new data source for public health 2013 [cited 2015 14-05-2015]. Available from: http://www.naccho.org/topics/infrastructure/informatics/resources/upload/dredze-naccho-webinar.pdf.
106. PewResearchCenter. Social media update 2014: Pew Research Center; 2015 [cited 2015 14-05-2015]. Available from: http://www.pewinternet.org/2015/01/09/social-media-update-2014/.
107. Dredze M. How social media will change public health. IEEE Intelligent Systems. 2012;27(4):81-4.
108. Paul M, Dredze M. You are what you tweet: analyzing Twitter for public health. International Conference on Weblogs and Social Media (ICWSM). 2011.
109. Smith M, Broniatowski D, Paul M, Dredze M. Tracking public awareness of influenza through Twitter. 3rd International Conference on Digital Disease Detection (DDD). 2015.
110. Paul M, Dredze M, Broniatowski D, Generous N. Worldwide influenza surveillance through Twitter. AAAI Workshop on the World Wide Web and Public Health Intelligence. 2015.
111. Coppersmith G, Dredze M, Harman C, Hollingshead K. From ADHD to SAD: analyzing the language of mental health on Twitter through self-reported diagnoses. NAACL Workshop on Computational Linguistics and Clinical Psychology. 2015.

112. Nakhasi A, Passarella R, Bell S, Paul M, Dredze M, Pronovost P. Malpractice and malcontent: analyzing medical complaints in Twitter. AAAI Fall Symposium on Information Retrieval and Knowledge Discovery in Biomedical Text. 2012.
113. Passarella R, Nakhasi A, Bell S, Paul M, Pronovost P, Dredze M. Twitter as a source for learning about patient safety events. Annual Symposium of the American Medical Informatics Association (AMIA). 2012.
114. Marmot M, Altman D, Cameron D, Dewar J, Thompson S, Wilcox M. The benefits and harms of breast cancer screening: an independent review. British Journal of Cancer. 2013;108:2205-40.
115. CDC. Breast cancer screening guidelines for women Atlanta: Centers for Disease Control and Prevention; 2016 [cited 2016 29-07-2016]. Available from: http://www.cdc.gov/cancer/breast/pdf/BreastCancerScreeningGuidelines.pdf.
116. Mai V, Sullivan T, Chiarelli A. Breast cancer screening program in Canada: successes and challenges. Salud Publica Mex. 2009;51(2):S228-S35.
117. Borugian M, Spinelli J, Abanto Z, Xu C, Wilkins R. Breast cancer incidence and neighbourhood income. Statistics Canada, 2011 Contract No.: 82-003-XPE.
118. Mahamoud A. Breast cancer screening in racialized women: implications for health equity. Wellesley Institute, 2014.
119. CDC. Cancer Screening - United States, 2010 2012 [cited 2015 16-05-2015]. Available from: http://www.cdc.gov/mmwr/preview/mmwrhtml/mm6103a1.htm.
120. Cruz-Castillo A, Hernández-Valero M, Hovick S, Campuzano-González M, Karam-Calderón M, Bustamante-Montes L. A study on the knowledge, perception, and use of breast cancer screening methods and quality of care among women from central Mexico. Journal of Cancer Education. 2014.
121. HealthTalkOnline. Reasons for not attending breast screening 2013 [cited 2015 06-08-2015]. Available from: http://www.healthtalk.org/peoples-experiences/cancer/breast-screening/reasons-not-attending-breast-screening.
122. University of Twente. Health belief model 2002 [cited 2017 18-01-2017]. Available from: https://www.utwente.nl/cw/theorieenoverzicht/Theory%20Clusters/Health%20Communication/Health_Belief_Model/.
123. Coldman A, Braun T, Gallagher R. The classification of ethnic status using name information. Journal Of Epidemiology And Community Health. 1988;42(2):390-5.
124. Choi B, Hanley A, Holowaty E, Dale D. Use of surnames to identify individuals of Chinese ancestry. American Journal of Epidemiology. 1993;138:723-34
125. Sheth T, Nair C, Nargundkar M, Anand S, Yusuf S. Cardiovascular and cancer mortality among Canadians of European, south Asian and Chinese origin from 1979 to 1993: an analysis of 1.2 million deaths. Canadian Medical Association Journal. 1999;161(2):132-8.
126. Shah B, Chiu M, Amin S, Ramani M, Sadry S, Tu J. Surname lists to identify South Asian and Chinese ethnicity from secondary data in Ontario, Canada: a validation study. BioMed Central - Medical Research Methodology. 2010;10(42):1-8.
127. Fiscella K, Fremont A. Use of geocoding and surname analysis to estimate race and ethnicity. Health Service Research. 2006;41:1482–500.
128. Gill P, Bhopal R, Kai J. Limitations and potential of country of birth as proxy for ethnic group. British Medical Journal. 2005;330(196).
129. Cambria E, Hussain A, Havasi C, Eckl C, Munro J, editors. Towards crowd validation of the UK National Health Service. Web Science Conference; 2010.
130. Greaves F, Ramirez-Cano D, Millett C, Darzi A, Donaldson L. Use of sentiment analysis for capturing patient experience from free-text comments posted online. Journal of Medical Internet Research. 2013;5(11):e239.

131. Statista. Number of monthly active Twitter users worldwide from 1st quarter 2010 to 1st quarter 2015 (in millions) 2015 [cited 2015 15-05-2015]. Available from: http://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/.
132. Tumasjan A, Sprenger T, Sandner P, Welpe I, editors. Predicting elections with Twitter: what 140 characters reveal about political sentiment Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media; 2010.
133. O'Connor B, Balasubramanyan R, Routledge B, Smith N, editors. From Tweets to polls: linking text sentiment to public opinion time series. Proceedings of the International AAAI Conference on Weblogs and Social Media; 2010; Washington, DC.
134. Bollen J, Mao H, Zeng X. Twitter mood predicts the stock market. Journal of Computational Science. 2011.
135. Cummings D, Oh H, Wang N. Who needs polls? Gauging public opinion from Twitter data. 2011.
136. Salathé M, Khandelwal S. Twitter & disease surveillance: a) crowdsourcing disease surveillance b) health behavior assessment: Center for Infectious Disease Dynamics; 2011 [cited 2015 15-05-2015]. Available from: http://www.syndromic.org/storage/documents/Twitter-Disease-Surveillance-salathe_ISDS.pdf.
137. Sakaki T, Okazaki M, Matsuo Y, editors. Earthquake shakes Twitter users: real-time event detection by social sensors. The World Wide Web Conference; 2010.
138. Robinson B, Power R, Cameron M. Disaster monitoring. Twitter: a digital socioscope2015.
139. Eysenbach G. Infodemiology and Infoveillance: framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the internet. Journal of Medical Internet Research. 2009;11(1):e11.
140. Salathé M, Khandelwal S. Assessing vaccination sentiments with online social media: implications for infectious disease dynamics and control. PLOS Computational Biology. 2011.
141. Myslín M, Zhu S, Chapman W, Conway M. Using Twitter to examine smoking behavior and perceptions of emerging tobacco products. Journal of Medical Internet Research. 2013;15(8):e174.
142. Chaudhry A, Glode L, Gillman M, Miller R. Trends in twitter use by physicians at the American society of clinical oncology annual meeting, 2010 and 2011. Journal of Oncology Practice. 2012;8(3):173–8.
143. Vance K, Howe W, Dellavalle R. Social internet sites as a source of public health information. Dermatologic Clinics. 2009;27:133-6.
144. Lapointe L, Ramaprasad J, Vedel I. Creating health awareness: a social media enabled collaboration. Health and Technology. 2014.
145. Sugawara Y, Narimatsu H, Hozawa A, Shao L, Otani K, Fukao A. Cancer patients on Twitter: a novel patient community on social media. BMC Research Notes. 2012;5:699.
146. Thackeray R, Burton S, Giraud-Carrier C, Rollins S, Draper C. Using Twitter for breast cancer prevention: an analysis of breast cancer awareness month. BMC Cancer. 2013;13:508.
147. Lyles C, López A, Pasick R, Sarkar U. "5 mins of uncomfyness is better than dealing with cancer 4 a lifetime": an exploratory qualitative analysis of cervical and breast cancer screening dialogue on Twitter. Journal of Cancer Education. 2013;28(1):127-33.
148. Canada LaA. 1901 Census: Government of Canada; 2016 [cited 2017 25-01-2017]. Available from: http://www.bac-lac.gc.ca/eng/census/1901/Pages/about-census.aspx.
149. FamilySearch. Canada Census, 1901. Library and Archives of Canada, Ottawa.
150. University of Victoria. 1901 database: Canadian families project 2006 [cited 2015 06-12-2014]. Available from: http://web.uvic.ca/hrd/cfp/data/index.html.
151. Automatedgenealogy. 1901 census of Canada indexing project 2009 [cited 2014 05-12-2014]. Available from: http://www.automatedgenealogy.com/census/index.jsp.

152. Library and Archives Canada. About the 1901 census: Library and Archives Canada; [cited 2015 16-05-2015]. Available from: http://www.bac-lac.gc.ca/eng/census/1901/Pages/about-census.aspx.
153. Wang W, Parker K, Passel J, Patten E, Motel S, Taylor P. The rise of intermarriage. Rates, characteristics vary by race and gender. PewResearchCenter. 2012.
154. Rehling J. Native American languages 1996 [cited 2017 25-01-2017]. Available from: http://www.cogsci.indiana.edu/farg/rehling/nativeAm/ling.html.
155. Mithun M. The languages of Native North America (Cambridge language surveys): Cambridge University Press; 2001.
156. Statistics Canada. Table 1 - Population with an Aboriginal mother tongue by language family, main languages within these families and their main provincial and territorial concentrations, Canada, 2011 2015 [cited 2017 25-01-2017].
157. Wikipedia. List of First Nations peoples 2006 [cited 2017 25-01-2017]. Available from: https://en.wikipedia.org/wiki/List_of_First_Nations_peoples.
158. Waldman C. Encyclopedia of Native American tribes (facts on file library of American history). 3rd ed: Checkmark Books; 2006.
159. Numpy. Numpy 2016 [cited 2017 25-01-2017]. Available from: http://www.numpy.org/.
160. SciKitLearn. SciKitLearn user guide 2014 [cited 2015 25-08-2015]. Available from: http://scikit-learn.org/stable/user_guide.html.
161. Pydata. Pandas - Python data analysis library 2016 [cited 2017 25-01-2017]. Available from: http://pandas.pydata.org/.
162. Wong K, Davis F, Zaïane O, Yutaka Y, editors. Sentiment Analysis of Breast Cancer Screening in the United States using Twitter. Proceedings of the 8th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management; 2016; Porto, Portugal.
163. Digital Insights. Social media 2014 statistics – an interactive infographic you've been waiting for 2014 [cited 2017 26-01-2017]. Available from: http://blog.digitalinsights.in/social-media-users-2014-stats-numbers/05205287.html.
164. Roesslein J. Tweepy documentation 2009 [cited 2015 25-08-2015]. Available from: http://tweepy.readthedocs.org/en/v3.2.0/.
165. Twitter. Geo guidelines: Twitter; 2014 [cited 2014 06-03-2014]. Available from: https://dev.twitter.com/overview/terms/geo-developer-guidelines.
166. MapQuest. Geocoding API 2014 [cited 2014 03-29-2014]. Available from: https://developer.mapquest.com/products/geocoding.
167. CDC-BRFSS. About BRFSS 2014 [cited 2015 25-08-2015]. Available from: http://www.cdc.gov/brfss/about/index.htm.
168. CDC-BRFSS. Annual survey data: CDC; 2015 [cited 2015 25-08-2015]. Available from: http://www.cdc.gov/brfss/annual_data/annual_data.htm.
169. Hutto C, Gilbert E. VADER: a parsimonious rule-based model for sentiment analysis of social media text. Association for the Advancement of Artificial Intelligence. 2014.
170. ArcGIS. How optimized hot spot analysis works: Environmental Systems Research Institute, Inc.; 2015 [cited 2015 06-11-2015]. Available from: http://desktop.arcgis.com/en/desktop/latest/tools/spatial-statistics-toolbox/how-optimized-hot-spot-analysis-works.htm.
171. Yudell M, Roberts D, DeSalle R, Tishkoff S. Taking race out of human genetics. Science. 2016;351(6273):564-5.
172. da Silva Santos D, Palomares N, Normando D, Quintão C. Race versus ethnicity: differing for better application Dental Press Journal of Orthodontics. 2010;15(3):121-4.
173. American Anthropological Association. Statement on race 1998 [cited 2015 05-16-15]. Available from: www.aaanet.org/stmts/racepp.htm.

174. Winant H. Race and race theory. Annual Review of Sociology. 2000.
175. Last J. A dictionary of epidemiology. 3rd ed. New York: Oxford University Press; 1995.
176. Riley J. Race and ethnicity: the biological versus the socialized 2003 [cited 2016 31-08-2016]. Available from: http://www.longwood.edu/media/diversity-and-inclusion/public-site/Race_and_Ethnicity.pdf.
177. Matsumoto-Gray K. Categorization: connections between language and society Language, Meaning, and Society 2009;2:107-35.
178. Bhopal R. Ethnicity, race, epidemiology, and public health. 6th ed. Detels R, Gulliford M, Karim Q, Tan C, editors: Oxford University Press: Oxford Textbook of Global Public Health; 2015.
179. Statistics Canada. Ethnic diversity and immigration: health status and access to health care 2016 [cited 2016 20-10-2016]. Available from: http://www5.statcan.gc.ca/subject-sujet/subtheme-soustheme?pid=30000&id=30003&lang=eng&more=0.
180. BARHII. A public health framework for reducing health inequities: Bay area regional health inequities initiative 2015 [cited 2017 21-03-2017]. Available from: http://barhii.org/framework/.
181. Lin S, Kelsey J. Use of race and ethnicity in epidemiologic research: concepts, methodological issues, and suggestions for research. Epidemiologic Reviews. 2000;22(2):187-202.
182. Canadian Public Health Association. An investment in public health: an investment in the public's health brief to the standing committee on finance 2008.
183. Canadian Institute for Health Information. Pan-Canadian dialogue to advance the measurement of equity in health care: proceedings report. 2016.
184. World Health Organization Europe. The social determinants of health: the solid facts. 2003.
185. Mikkonen J, Raphael D. Social determinants of health: the Canadian facts. York University, 2010.
186. Ahmed S, Shahid R. Disparity in cancer care: a Canadian perspective. Current Oncology. 2012;19(6).
187. Liu R, So L, Mohan S, Khan N, King K, Quan H. Cardiovascular risk factors in ethnic populations within Canada: results from national cross-sectional surveys. Open Medicine. 2010;4(3):e143-53.
188. Del Amo J, Jarrin I, May M, Dabis F, Crane H, Podzamczer D, et al. Influence of geographical origin and ethnicity on mortality in patients on antiretroviral therapy in Canada, Europe, and the United States. Clinical Infectious Diseases. 2013;56(12):1800-9.
189. Poureslami I, Rootman I, Balka E, Devarakonda R, Hatch J, Fitzgerald J. A systematic review of asthma and health literacy: a cultural-ethnic perspective in Canada. Medscape General Medicine. 2007;9(3):40.
190. Rosenberg M, Lo L. Ethnicity and utilization of family physicians: a case study of Mainland Chinese immigrants in Toronto, Canada. Social Science and Medicine. 2008;67(9):1410-22.
191. Prasad G. Renal transplantation for ethnic minorities in Canada: inequity in access and outcomes? Kidney International. 2007;72(4):390-2.
192. Agency for Healthcare Research and Quality. Disparities in healthcare quality among racial and ethnic groups: selected findings from the 2011 National Healthcare Quality and Disparities Reports. Fact Sheet. Rockville, MD: 2012 Contract No.: Publication No. 12-0006-1-EF.
193. Healthy People Team. Disparities: Office of Disease Prevention and Health Promotion; 2010 [cited 2015 02-24-2015]. Available from: http://www.healthypeople.gov/2020/about/foundation-health-measures/Disparities.
194. Alliance for Health Reform. Closing the gap: racial and ethnic disparities in healthcare. Journal of the National Medical Association. 2004;96(4):436-40.
195. Statistics Canada. Canada's ethnocultural mosaic, 2006 census. . 2008 Contract No.: Catalogue no. 97-562-X.
196. Fisher M. A revealing map of the world's most and least ethnically diverse countries: The Washington Post; 2013 [03-05-2015]. Available from:

http://www.washingtonpost.com/blogs/worldviews/wp/2013/05/16/a-revealing-map-of-the-worlds-most-and-least-ethnically-diverse-countries/.
197. Statistics Canada. 2006 census: Aboriginal peoples in Canada in 2006: Inuit, Métis and First Nations, 2006 census: highlights 2009 [cited 2016 17-10-2016]. Available from: https://www12.statcan.gc.ca/census-recensement/2006/as-sa/97-558/p1-eng.cfm.
198. Statistics Canada. National Population Health Survey: Household Component, Longitudinal (NPHS): detailed information for 2010-2011 (Cycle 9) 2012 [cited 2016 18-10-2016]. Available from: http://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=3225.
199. Statistics Canada. Ethnic Diversity Survey (EDS): detailed information for 2002 2007 [cited 2016 18-10-2016]. Available from: http://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=4508.
200. Statistics Canada. National Longitudinal Survey of Children and Youth (NLSCY): detailed information for 2008-2009 (Cycle 8) 2010 [cited 2016 18-10-2016]. Available from: http://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=4450.
201. Statistics Canada. Aboriginal Children's Survey (ACS): detailed information for 2006 2007 [cited 2016 18-10-2016]. Available from: http://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=5108.
202. Centre for Education Statistics. Federal data scan: Aboriginal data in Statistics Canada's education data sources. . 2009.
203. Canadian Institute for Health Information. Canadian Patient Experiences Survey—Inpatient Care Procedure Manual, May 2014. 2014.
204. Institute for Clinical Evaluative Sciences. Datad dictionary - CORR 2016 [cited 2016 29-11-2016]. Available from: https://datadictionary.ices.on.ca/Applications/DataDictionary/Variables.aspx?LibName=CORR&MemName=DONOR&Variable=RACIAL_ORIGIN_CODE.
205. Young S, Nishri E, Candido E, Marrett L. Colorectal cancer incidence in the Aboriginal population of Ontario, 1998 to 2009. Statistics Canada, 2015.
206. National Research Council. Eliminating health disparities: measurement and data needs. Washington, DC: Committee on National Statistics, Division of Behavioral and Social Sciences and Education, 2004.
207. Moubarac J. Persisting problems related to race and ethnicity in public health and epidemiology research. Rev Saúde Pública. 2013;47(1):104-15.
208. Nestel S. The impact of race and racism on Canadians' health. Wellesley Institute, 2012.
209. Randall V. Eliminating racial discirmination in health care. Totowa, NJ: Humana Press; 2007.
210. Kendrick J, Nuccio E, Leiferman A, Sauaia A. Primary care providers perceptions of racial/ethnic and socioeconomic disparities in hypertension control. American Journal of Hypertension. 2015;28(9):1091-7.
211. Iqbal G, Johnson M, Szczepura A, Gumber A, Wilson S, Dunn J. Ethnicity data collection in the UK: the healthcare professional's perspective. Diversity and Equality in Health and Care. 2012;9:281-90.
212. Tri-Hospital. We ask because we care: the Tri-Hospital + TPH health equity data collection research project report. Toronto: Toronto Public Health, St. Michael's Hospital, CAMH, and Mount Sinai Hospital, 2013.
213. Council of Canadian Academies. Accessing health and health-related data in Canada. Ottawa: The Council of Canadian Academies, 2015.
214. Loppie S, Reading C, de Leeuw S. Aboriginal experiences with racism and its impacts. Prince Geoge, British Columbia: National Collaborating Centre for Aboriginal Health, 2014.
215. Bahdi R, Parsons O, Sandborn T. Racial profiling position paper. BC Civil Liberties Association, 2009.

216. Ontario Human Rights Commission. Count me in! Collecting human rights-based data. Toronto, Ontario: Government of Ontario, 2010.
217. Canadian Institute for Health Research. CIHR best practices for protecting privacy in health research. Ottawa, Ontario: 2005.
218. Statistics Canada. 2011 census of population: final response rates 2015 [cited 2016 17-10-2016]. Available from: https://www12.statcan.gc.ca/census-recensement/2011/ref/about-apropos/rates-taux-eng.cfm.
219. Peel Public Health. About census data: Region of Peel; 2012 [cited 2016 17-10-2016]. Available from: https://www.peelregion.ca/health/statusdata/DataSources/HSD12_11.asp.
220. Statistics Canada. National Household Survey: final response rates 2015 [cited 2016 18-10-2016]. Available from: https://www12.statcan.gc.ca/nhs-enm/2011/ref/about-apropos/nhs-enm_r012.cfm?Lang=E.
221. Statistics Canada. Canadian Community Health Survey - mental health (CCHS) 2013 [cited 2016 17-10-2016]. Available from: http://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&Id=119789.
222. Statistics Canada. Canadian Community Health Survey (CCHS) annual component. 2011.
223. Peel Public health. About the Canadian Community Health Survey 2013 [cited 2016 17-10-2016]. Available from: https://www.peelregion.ca/health/statusdata/DataSources/HSD12_2.asp.
224. Statistics Canada. Data sources, methods, and limitations: Canadian Community Health Survey 2015 [cited 2016 17-10-2016]. Available from: http://www.statcan.gc.ca/pub/82-622-x/2011007/method-eng.htm.
225. Statistics Canada. Canadian Health Measures Survey (CHMS): detailed information for January 2014 to December 2015 (cycle 4) 2014 [cited 2016 18-10-2016].
226. Statistics Canada. Aboriginal Peoples Survey (APS): detailed information for 2012. 2012. Available from: http://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=3250&lang=en&db=imdb&adm=8&dis=2.
227. Eller J. Ethnicity, culture, and "the past". MichiganQuarterlyReview. 1997;36(4).
228. Sundquist J, Johansson S. The influence of socioeconomic status, ethincity and lifestyle on body mass index in a longitudinal study. International Epidemiological Association. 1998;27:57-63.
229. da Silva Santos D, Palomares N, Normando D, Quintão C. Race versus ethnicity: differing for better application. Dental Press Journal of Orthodontics. 2010;15(3):121-4.
230. Fearon J. Ethnic and cultural diversity by country. Journal of Economic Growth. 2003;8(2):195-222.
231. National Collaborating Centre for Aboriginal Health. An overview of Aboriginal health in Canada. 2013.
232. Consumer Financial Protection Bureau. Using publicly available information to proxy for unidentified race and ethnicity: a methodology and assessment. 2014.
233. Elliott M, Morrison P, Fremont A, McCaffrey D, Pantoja P, Lurie N. Using the Census Bureau's surname list to improve estimates of race/ethnicity and associated disparities. Health Services and Outcomes Research Methodology. 2009;9(69).
234. Library and Archives Canada. 1871 Census (Canada) 2017 [cited 2017 21-03-2017]. Available from: http://www.bac-lac.gc.ca/eng/census/1871/Pages/about-census.aspx.
235. Indigenous and Northern Affairs Canada. First Nations in Canada: Government of Canada; 2013 [cited 2017 21-03-2017]. Available from: https://www.aadnc-aandc.gc.ca/eng/1307460755710/1307460872523.
236. Christian Aboriginal Infrastructure Developments. Meaningful consultation in Canada: the alternative to forced Aboriginal assimilation. 2009.

237. Joseph B. The Indian Act naming policies: Indigenous Corporate Training Inc.; 2014 [cited 2017 21-03-2017]. Available from: http://www.ictinc.ca/indian-act-naming-policies.
238. Ng A. CS229 lecture notes (on classification and logistic regression). 2016.
239. Guestrin C. Learning logistic regressors by gradient descent. Machine Learning – CSE4462013.
240. Scikit-Learn. Linear model: logistic regression 2010 [cited 2014 04-06-2014]. Available from: http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html.
241. Smith B. Lagrange multipliers tutorial in the context of support vector machines. In: Newfoundland MUo, editor. St. John's, Newfoundland, Canada2004.
242. Scikit-Learn. 1.4. Support Vector Machines 2014 [cited 2016 27-11-2016]. Available from: http://scikit-learn.org/stable/modules/svm.html.
243. Ramanan D, Park D, Nikamanon P. ICS 273A: Machine Learning: lecture 11.1 Soft Margin SVM. 2008.
244. Scikit-Learn. 1.10. Decision Trees 2010 [cited 2016 30-10-2016]. Available from: http://scikit-learn.org/stable/modules/tree.html.
245. Tan P, Steinbach M, Kumar V. Chapter 4 classification: basic concepts, decision trees, and model evaluation. Introduction to data mining Pearson; 2005.
246. Statistics Canada. Aboriginal Peoples in Canada: First Nations People, Métis and Inuit Canada: Social and Aboriginal Statistics Division; 2015 [cited 2016 24-07-2016]. Available from: https://www12.statcan.gc.ca/nhs-enm/2011/as-sa/99-011-x/99-011-x2011001-eng.cfm.
247. Brownlee J. Classification accuracy is not enough: more performance measures you can use 2014 [cited 2016 30-10-2016].
248. Fawcett T. An introduction to ROC analysis. Pattern Recognition Letters. 2006;27:861–74.
249. Statistics Canada. Where do Aboriginal people live? 2007 [cited 2016 27-11-2016]. Available from: http://www41.statcan.gc.ca/2007/10000/ceb10000_003-eng.htm.
250. Northwest Territories Health and Social Services. Cancer in the Northwest Territories, 2001-2010. Northwest Territories: Government of the NWT, 2014.
251. Council of Ontario Universities. Aboriginal self-identification project final report. Toronto, Ontario: COU, 2013.
252. University of British Columbia. Aboriginal self-identification: University of British Columbia; 2015 [cited 2015 06-11-2015]. Available from: http://aboriginal.ubc.ca/students/aboriginal-self-identification/.
253. Khan M, Kobayashi K, Lee S, Vang Z. (In)visible minorities in Canadian health data and research. Population Change and Lifecourse Strategic Knowledge Cluster Discussion Paper Series/ Un Réseau stratégique de connaissances Changements de population et parcours de vie Document de travail. 2015;3(1):Article 5.
254. American Cancer Society. Cancer facts and figures 2015. Atlanta: American Cancer Society; 2015.
255. Myers E, Moorman P, Gierisch J, Havrilesky L, Grimm L, Ghate S, et al. Benefits and harms of breast cancer screening: a systematic review. JAMA. 2015;314(15):1615-34.
256. Janz N, Becker M. The health belief model: a decade later. Health Education Quarterly. 1984;11(1):1-47.
257. Fulton J, Buechner J, Scott H, DeBuono B, Feldman J, Smith R, et al. A study guided by the health belief model of the predictors of breast cancer screening of women ages 40 and older. Public Health Reports. 1991;106(4):410-20.
258. Wang W, Hsu S, Wang J, Huang L, Hsu W. Survey of breast cancer mammography screening behaviors in Eastern Taiwan based on a health belief model. Kaohsiung Journal of Medical Sciences. 2014;30:422-7.

259. Austin L, Ahmad F, McNally M, Stewart D. Breast and cervical cancer screening in Hispanic women: a literature review using the health belief model. Women's Health Issues. 2002;12(3):122-8.
260. Bryson E, Schafer E, Salizzoni E, Cosgrove A, Favaro D, Dawson R. Is perception reality? Identifying community health needs when perceptions of health do not align with public health and clinical data. SM Journal of Community Medicine. 2016;2(1):1013.
261. Pang B, Lee L. 4.1.2 Subjectivity detection and opinion identification. Opinion mining and sentiment analysis: Now Publishers Inc.; 2008.
262. Kumar S, Morstatter F, Liu H. Twitter data analytics: Springer; 2013.
263. Centers for Disease Control and Prevention. Behavioral risk factor surveillance system Atlanta, GA: CDC; 2016 [cited 2015 05-12-2015]. Available from: http://www.cdc.gov/brfss/.
264. MapOfUSA. US population density map 2007 [cited 2015 16-12-2015]. Available from: http://www.mapofusa.net/us-population-density-map.htm.
265. Mitra T, Counts S, Pennebaker J. Understanding anti-vaccination attitudes in social media. Tenth International AAAI Conference on Web and Social Media: AAAI; 2016.
266. Brooks B. Using Twitter data to identify geographic clustering of anti-vaccination sentiments. Seattle: University of Washington; 2014.
267. Pulman S. Multi-dimensional sentiment analysis Oxford: Dept. of Computer Science, Oxford University; 2014 [cited 2016 21-08-2016]. Available from: http://www.lt-innovate.org/sites/default/files/lt_accelerate_files/13.30%20Stephen_Pulman_UNIV_OXFORD.pdf.
268. Khan M, Kobayashi K, Lee S, Vang Z. (In)visible minorities in Canadian health data and research. 2015.
269. Pulver L, Haswell M, Ring I, Waldon J, Clark W, Whetung V, et al. Indigenous health – Australia, Canada, Aotearoa New Zealand and the United States - laying claim to a future that embraces health for us all. World Health Organzation, 2010.
270. Centers for Disease Control and Prevention. Behavioral Risk Factor Surveillance System: BRFSS questionnaires 2016 [cited 2017 19-01-2017]. Available from: https://www.cdc.gov/brfss/questionnaires/index.htm.
271. Mislove A, Lehmann S, Ahn Y, Onnela J, Rosenquist N, editors. Understanding the demographics of Twitter users. Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM'11); 2011; Barcelona, Spain.
272. Greenwood S, Perrin A, Duggan M. Social media update 2016: Pew Research Center; 2016 [cited 2017 31-01-2017].
273. Culotta A, editor Reducing sampling bias in social media data for county health inference. JSM Proceedings; 2014.
274. Lucas R, McMichael A. Association or causation: evaluating links between "environment and disease". Bulletin of the World Health Organization. 2005;83(10):792-5.
275. McCord M, Chuah M, editors. Spam detection on Twitter using traditional classifiers. 8th International Conference, ATC 2011; 2011; Banff, Canada.
276. Hutto C, Gilbert E. VaderSentiment 2.5: Pypi; 2016 [cited 2017 31-01-2017]. Available from: https://pypi.python.org/pypi/vaderSentiment.
277. Asghar M, Ahmad A, Qasim M, Zahra S, Kundi F. SentiHealth: creating health-related sentiment lexicon using hybrid approach. Springerplus. 2016;5(1):1139.
278. Goeuriot L, Na J, Kyaing W, Khoo C, Chang Y, Theng Y, et al., editors. Sentiment lexicons for health-related opinion mining. IHI'12 Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium; 2012; Miami, Florida, USA.
279. Kim S, Hovy E, editors. Identifying and analyzing judgment opinions. Proceedings of the Human Language Technology/North American Association of Computational Linguistics conference (HLT-NAACL 2006); 2006; New York, United States.

280. Pulman S. Multi-dimensional sentiment analysis: Department of Computer Science, Oxford University; 2014 [cited 2017 31-01-2017]. Available from: http://www.lt-innovate.org/sites/default/files/lt_accelerate_files/13.30%20Stephen_Pulman_UNIV_OXFORD.pdf.
281. Thurau T, Wiertz C, Feldhaus F. Exploring the 'Twitter effect:' an investigation of the impact of microblogging word of mouth on consumers' early adoption of new products. SSRN Electronic Journal. 2012.