Emerging Uses of Big Data in Immigration Research

Final Report Submitted to SSHRC

October 13th, 2016

Grant No. 421-2015-2036

William Ashton, Pallabi Bhattacharyya, Eleni Galatsanou, Sally Ogoe and Lori Wilkinson[1]

This research was supported by the

Social Sciences and Humanities Research Council of Canada.

[1] The names are written in alphabetical order and do not reflect order of authorship. Special thanks to Kim Lemky for her input with proposal development.

Table of Contents

Key messages

How has big data been used in immigration studies.

- There is an increasing trend in utilizing big data in immigration studies from 2007 onwards, with more and more published studies using big data than ever before. The majority of studies are published in journal articles while Labor market and Social Integration are the two main research themes that dominate the literature.
- Close to 60% of immigration studies use Statistics Canada data for their research, either in the form of the Census, Longitudinal and other surveys but also linked by Statistics Canada databases. The Canadian Census and the Longitudinal Survey of Immigrants to Canada (LSIC) are utilized the most among the large databases.

Challenges with using big data in immigration studies

- Big data sets, despite their large number of participants and variables, have issues of population underrepresentation, especially in administrative data and often contribute to inaccurate and unsettled relationship between selected variables. Large data sets often lack specific variables or additional explanation on variables and immigration researchers rely on assumptions and proxies to complete their studies.
- Methodological challenges and challenges with data manipulation and validation influences the kind of studies researchers can do and the kind of questions they can ask. Storage, management, and sharing of big data remain a challenge for all users of big data. Without a university-affiliation, most of this data is out of the reach of non-government settlement organizations, groups that can critically analyze this information for practical purposes.
- There are ethical issues with the use of linked data and with administrative data and many of the large datasets did not seek ethical approval by participants. Credibility of results is a major issue since data from administrative sources, collected from social media, internet and online communication sources are subject to bias.

Opportunities

- Big data can prove very beneficial for settlement and community organizations for designing and implementation of policies and programs, provided it is timely released and easy to access. Researchers should be encouraged to partner with institutions and organizations to examine questions relevant to the practical needs of the settlement sector.
- Great knowledge can be gained from examining data from online sources (large datasets, social media, web login history, IP addresses and others) and can provide new opportunities for immigration studies. Addressing inherent risks and train more researchers in statistical modeling and data analysis are some mandatory steps in order for immigration research to utilize all forms of big data.
- By the year 2018, new developments concerning linked administrative and survey data are expected to be released and will give additional opportunities for further research on a variety of different topics.

Executive summary

This project seeks to understand the use, challenges, and opportunities of big data sets in immigration research. Big data are large sets of data whose size is beyond the ability of typical software tools to capture, store, manage, and analyze. The data sets come from government surveys, from geolocation like Facebook tags, and digital transaction like financial services. At the outset of this project, it is not clear how or if big data are being used by researchers to dig deeper and more comprehensively about the lives of immigrants. Such findings can inform immigration settlement services, business products and services offered to immigrants, along with government policies and programs.

A scoping review method isolated key themes embedded within a large body of immigration literature. The establish six-stage framework by Levac, Colquuhoum, and O'Brien (2010) enabled a rigorous approach that identified 453 papers, and after confirming high levels of inter-rater reliability, there were 251 immigration studies (1994-2016) used in this knowledge synthesis project. About 60% were journal articles, with another 20% as reports, and the balance from grey literature and theses. The papers examined five themes, yet dominated by labour market (57%) and social integration (25%) papers. The studies used 15 data bases, with 60% of them drawing from Statistics Canada and 65% used one data set and 15% utilized more than three sets.

Authors reported many challenges that were grouped under five topics, including sampling issues, low response rates, invalidated variables, access, and confidentiality. The authors identified opportunities as well, ranging from the need for diverse knowledge on data manipulation given the complexity and growing nature of some data sets, to design experiments on unprecedented scale, skills development of researchers to use data sets, linking data bases to examine different aspects of immigrants' lives, and need to advance hardware and software along with tools and techniques of data mining.

Big data has great potential in immigration research. As mentioned earlier, geocoded data can help aid organizations identify where people in need might be located, but also what they might need! The Syrian Humanitarian Tracker has been used by thousands of internally displaced people and asylum seekers abroad to identify safe travel routes and to avoid human traffickers. This kind of data has been used in North America to provide real time alerts, with information, about missing children (e.g., Amber Alerts in Canada). The monitoring of social media data by non-government aid agencies can help them better determine how many people might be waiting to access English or French language courses.

Canada has recently provided extensive and almost on-demand data on the Syrian refugee arrivals. Through Immigration Refugee and Citizenship Canada's online database, the public can now view maps which are uploaded from their website and provide geographically linked information on the number and location of resettled Syrian refugees in Canada. The government of Canada's "Open Data" initiative provides some data on citizenship uptake, international students, temporary foreign work permit holders and other immigration related issues. These are static tables, but provide additional information for users which were not previously made available. Statistics Canada also provides tabular information based on results from the National Census that allow site visitors to identify various trends in immigrants and labour market identifiers, ethnic identity and second-generation and language use among newcomers.

Statistics Canada, together with Immigration Refugees and Citizenship Canada, have been working to provide researchers with even more data opportunities. This is in addition to the wealth of data Statistics Canada provides in both Public Use Microdata files which novice and intermediate statisticians can use to produce their own statistical estimates and equations (e.g., Ethnic Diversity Survey, Longitudinal Survey of Immigrants to Canada, etc.). More advanced data users affiliated with universities can access the master data files of over three dozen current surveys. This allows users to look at small-scale trends and patterns in data that cannot be released to the general public. Although it is possible for independent researchers to gain access to administrative-level data such as the master data file for IMDB which is housed in Ottawa, these opportunities are given to only a few select researchers. There have been some very innovative and recent uses of the data, such as the GIS Mapping Project (Garcea et al., 2016) which uses data from the Immigrant Landing File and IMDB to geo-code landings and other records among immigrants and refugees to Canada's western and northern regions.

1. Introduction

What do big data and refugee movements have in common? Turns out, a lot actually. The past 18 months have been a boon in the speed, volume and type of data being shared both publicly and by governments concerning refugee movements in Europe, Africa and the Middle East. Large-scale migration from Syria, although over five years old, really began in earnest in Spring 2015 when large numbers of asylum seekers braved the 4KM journey between Bodrum Turkey and Kos Greece in rickety plastic dinghies and broken down overcrowded fishing boats. The signing of the *EU-Turkey Repatriation of Refugees Agreement* in late 2015 was supposed to end, or at least greatly decrease, the number of would-be refugees making the perilous journey across the Mediterranean Ocean. Instead, it is estimated that by early December 2016, over 170,000 new asylum seekers will arrive on the shores of Italy, surpassing the numbers arriving in Greece (IOM, 2016). How do we know this? The IOM releases weekly figures of arrivals, deportations and resettlement of refugees in Europe. They depend on data sharing agreements with the various countries and then tabulate and distribute this information to a network of policy analysts, researchers and non-governmental organizations on a weekly basis. The Canadian government did the same, reporting weekly on Immigration Refugee and Citizenship Canada's (IRCC) website on the numbers and locations of the newly arrived Syrians to Canada. It is the first time in history that this level of detailed, geographically connected data was widely distributed and publicly available in such a short period of time. Both the IOM and IRCC report that traffic to these websites has increased dramatically in the past year as the thirst for knowledge and data about the newly arriving refugees increased.

These refugee movements, along with the various government and non-government interventions to count and account for refugee movement, got our team thinking about the extent to which big data has been used in immigration research in Canada. Our project aims to identify studies using big data and to offer some insight to the type of data they use, the topics they pursue and where these studies might be published. The movement to use 'big data' in all sorts of research and decision making is part of a growing trend that takes advantage of the large amount of data we generate as humans in the 21st century. In 2015, "we produced as much data as was created in all previous years of human civilization" (Ratti and Helbing, 2016: 1). This data has been used to harness the movement and activities of the Syrian people as they flee the conflict that mires their country. The Syria Tracker Crisis Map (Humanitarian Tracker, 2016), is an app created by displaced

Syrians intended to provide information for Syrians on the run (Rathnam, 2015). The data comes largely from social media platforms such as Twitter, Facebook and Skype, along with media reports to produce information which is posted to the site and intended to assist NGOs and other organizations with the needs of displaced Syrians. It has helped identify geographic regions where disease amongst the displaced people might occur next, assisted in the prevention of human trafficking, and is credited with predicting when the next attack will occur (Rathnam, 2015). We wondered if these initiatives were taking place in the Canadian research context.

This report begins with a discussion of the context of big data and research in general. The implications of using this kind of data, along with the methodology of the study follow. Section four details the quantitative findings of our study, while sections 5 and 6 detail the knowledge mobilization plans and some concluding comments.

2. Context – the issue

Traditionally, big data has been defined as mega-sized sets of information that are too large for common data processing software to handle. Not so long ago, big data would have included the public use microdata files (PUMF) of the Census of Canada. Back in the mid-1990s, micro computer power was so under-developed that PUMF files had to be hand-loaded onto a microprocessor and time to analyze the information had to be 'booked' on an institution's mainframe computer. In that age, a census file of half a million people would take hours to compute even the simplest cross-tabulations because computer processing power was weak. By the start of the 21st century, microprocessors became smaller and faster and our ability to store data digitally, instead of by analog, has meant an explosion in data storing, sharing and computing capabilities. And because the cost of owning a computer and storing large amounts of data have greatly declined, almost anyone with database and statistical knowledge can work with large datasets. Now, researchers can download and analyze census and other similar data on their own and can manipulate the data in mere seconds to produce similar (and often better) and faster results than in previous years.

Alongside the development of better personal computer processors, companies, governments and other organizations began to discover that valid predictions and trends could be calculated by examining the data they already collect. And we collect a lot of data. It's been estimated that 2.5 exabytes of new data are generated worldwide on a daily basis (Hilbert and López, 2011). This, in addition to the data collecting power of various software programs developed to glean information from the internet has meant that it has become both easier than ever to conduct data analyses, but also meant a proliferation in data collection activities including social media and internet traffic monitoring, administrative data mining, and linking data between datasets, the extent of which many of these activities the public are unaware. Data isn't just limited to the affluent or in developed nations. Nearly 3.5 billion people have some access to the Internet, representing 40% of the world's population and almost half of them are currently living in Asia (Internet Live Statistics, 2016). According to ICT (2016) there are over 7 billion mobile phone subscriptions representing about 60% of the world's population. This means a growing number of people, regardless of their living situation, have some access to electronic communication, a type of data that is easily collected and monitored. This type of data is sometimes mined in big data projects so our ability to measure certain features of life and certainly aspects of online connectivity have greatly contributed to the big data movement. This data is particularly important in terms of adding a geographic dimension—where people are making inquiries about information is a very important indicator for several social and health problems. Internet searches for 'flu' help health researchers identify potential outbreaks and clusters before they are reported to health authorities. The Syrian Humanitarian Tracker mentioned above is another example of geo-coded data used.

Governments are increasingly using their administrative databases to track trends among its citizens and to monitor certain aspects of their lives. This administrative data is a different type of information, no less 'messy' than the data mined from internet sources. Government tax records, provincial health databases and criminal justice administrative data have all been used to produce reports, inform policy and provide meaningful, large-scale information to government offices. Historically, this data has only been available to government employees. Outsiders, such as NGOs, independent researchers and academics were largely prevented from accessing these interesting data sources due to privacy concerns. Today, however, governments are increasingly connecting with academics and universities to analyze these data. Two examples are the Manitoba Centre for Health Policy Research and Policywise for Children and Families in Alberta are two initiatives where independent researchers routinely gain access to various administrative databases with the express purpose of producing analyses and reports that inform government policy and decision making. This type of data is also associated with the big data movement because of the size and ever changing nature of the information held within the database.

Less common are other, usually a bit smaller, forms of big data that rely on data linkage. Data linkage refers to the connection of two or more unrelated sets of data. In the immigration field, the most widely used and well known linked data is the IMDB or Immigrant Database. It links the Immigrant Landing File and the Canadian Tax Files for all immigrants who have landed in Canada since 1980. The data is prepared as a 'cube' meaning that it is less flexible than other data as users are usually constrained to changing the conditions of three or four variables at a time and that statistical analysis is not largely possible. We can think of these kinds of data as more static. In an attempt to make them more user friendly, administrators have mounted them in platforms such as 2020 Ivision Browser which allow users to create their own tables by manipulating a series of rows and columns. Sadly, these formats tend to be clumsy and frustrating for novice users. Additionally, these large data bases are sources of potential identity disclosure, so only a few vetted and pre-approved researchers and government employees ever have a chance to use these very complex but rich data.

It is this latest form of data, linked databases that have produced the most recent and voluminous immigration research in Canada. "Big data refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze" (Manyika et al., 2011: 1). But what is typical database software? Although many well-known and commonly used statistical packages do have limits as to the number of gigabytes or terabytes they can process, their computing limitations don't seem to prevent researchers and statisticians from harnessing the vast information contained within them.

The OECD (2013) has characterized big data as having these three attributes:

- Volume: the amount of data within a database "challenges the capacities of traditional software and storage"

- Velocity: data can be distributed, disseminated and analyzed with increasing speed. The goal is to process data in real time
- Variety: data that comes from a variety of different data sources, preferably triangulated

There are other 'Vs' that have been added by various researchers including value, which refers to the economic and social uses of data (Laczko and Rango, 2014). Other Vs identified include value, visualization, veracity and variability (van Rijmenam, 2014). As McNulty (2014) observes, it is the 3 Vs that remain the most widely used to categorize data and are the categories we rely upon in this report.

3. Implications and Potential Uses of big data in Immigration Research

Big data has great potential in immigration research. As mentioned earlier, geocoded data can help aid organizations identify where people in need might be located, but also what they might need! The Syrian Humanitarian Tracker has been used by thousands of internally displaced people and asylum seekers abroad to identify safe travel routes and to avoid human traffickers. This kind of data has been used in North America to provide real time alerts, with information, about missing children (e.g., Amber Alerts in Canada). The monitoring of social media data by non-government aid agencies can help them better determine how many people might be waiting to access English or French language courses.

Canada has recently provided extensive and almost on-demand data on the Syrian refugee arrivals. Through Immigration Refugee and Citizenship Canada's online database, the public can now view maps which are uploaded from their website and provide geographically linked information on the number and location of resettled Syrian refugees in Canada. The government of Canada's "Open Data" initiative provides some data on citizenship uptake, international students, temporary foreign work permit holders and other immigration related issues. These are static tables, but provide additional information for users which were not previously made available. Statistics Canada also provides tabular information based on results from the National Census that allow site visitors to identify various trends in immigrants and labour market identifiers, ethnic identity and second-generation and language use among newcomers.

Statistics Canada, together with Immigration Refugees and Citizenship Canada, have been working to provide researchers with even more data opportunities. This is in addition to the wealth of data Statistics Canada provides in both Public Use Microdata files which novice and intermediate statisticians can use to produce their own statistical estimates and equations (e.g., Ethnic Diversity Survey, Longitudinal Survey of Immigrants to Canada, etc.). More advanced data users affiliated with universities can access the master data files of over three dozen current surveys. This allows users to look at small-scale trends and patterns in data that cannot be released to the general public. Although it is possible for independent researchers to gain access to administrative-level data such as the master data file for IMDB which is housed in Ottawa, these opportunities are given to only a few select researchers. There have been some very innovative and recent uses of the data, such as the GIS Mapping Project (Garcea et al., 2016) which uses data from the Immigrant Landing File and IMDB to geo-code landings and other records among immigrants and refugees to Canada's western and northern regions.

In summary, there are many possibilities for using this data and non-government agencies, researchers and policy analysts are making increasing use of this timely and important data. This report outlines some of the outcomes of this research.

4. Methodology

This project seeks to understand the use and challenges of big data in immigration research. A scoping review method served to isolate key themes embedded within subject areas which enabled a preliminary summary of a large body of immigration literature. Given a short timeline, Levac, Colquhoun and O'Brien's (2010) six-stage framework provided a rigorous method for mapping literature.

Step one of the scoping review is determining the search terms to be used and in our case, the databases most commonly used. We used several search terms including: administrative data, refugees, mental health, evidence based research, data driven research, big data immigration, labour market income and jobs, among others. We also searched for major databases which we were aware of. These include: big data, Longitudinal Survey of Immigrants to Canada (LSIC), Landed Immigrants Data file (LIDS), Immigrant Landing File (ILF), Immigrant Database (IMDB), Census of Canada, National Household Survey (NHS), and Canadian Community Health Survey (CCHS).

The second stage identified articles from journals as well as other relevant non-indexed literature including government reports, OECD reports, and unpublished studies. These search terms were used to search university library websites and databases, such as EBSCOhost that indexes and abstracts over 3,000 periodicals, and full text of over 1,500 periodicals. Many of these periodical journals are peer-reviewed. In addition, librarians from the University of Manitoba and Brandon University reviewed and their suggestions were used to assist us in the search procedures. This search resulted in a data pool with the relevant literature extracted.

The third stage of the study involved two researchers independently reviewing the abstracts, research themes, methodology and study limitations for each article and report. All this information helped in deciding if the item involved big data and how they have been used in immigration studies. The mapping of the data was done in two separate spreadsheets for a total of 453 studies. Both reviewers independently extracted the required information and coded the literature as "include", "exclude" or "uncertain". Twenty-four studies were duplicates of one another and 127 studies were excluded. A total of 251 studies eventually met the project criteria: being about a topic of immigration in Canada and possibly involving 'big data'. There were also about 60 general research papers and research articles that examine big data as a research methodology, which were separately coded for answering few of the study questions on challenges and opportunities.

The fourth stage involved confirming consistency of coding the articles between the two researchers. Ten random entries were given to each research assistant to recode as a process of checking validity and reliability in coding of data (for a total of 20 studies recoded). After small changes were made to ensure consistency they proceeded to code all articles. A third researcher scrutinized the coding results and resolved any discrepancies. These checks served to strengthen the reliability of the coding.

The fifth stage involved collating and summarizing results with a descriptive numeric analysis (number of documents, type, year published, etc.) and a thematic analysis (e.g. identify themes used). Reporting results included identifying implications for research, policy and practice.

The last stage, consultation, was completed with the advisory panel for the purpose of knowledge mobilization. There were in between consultations with the advisory panel for including their ideas and sharing the research process and outcomes from time to time during the whole tenure of the project.

There were few methodological challenges which the researchers faced while doing the scoping review. It was difficult to decide at the beginning if the articles in the search list were appropriate for the intended research topic, especially if the data that was used was really 'big data'. In the end, we decided to code the data sets in a way that allowed us to easily identify the administrative linked data and could separate those from other types of data (some of which is arguably not 'big data' at all). We found that key words that authors used to describe their studies did not accurately reflect the actual content of the article. Our search was conducted using English terms only, which biased our search results. Given the six month time constraint to complete this project, we ran out of time to include the valuable French language papers. This way we might have missed some scoping reviews in the available literature. It was also very difficult to decide if we have included all the relevant articles and the only way we could assure that we have searched exhaustively, was through repeated search leading to similar articles showing up again and again. One of the principle investigators assisted the team by identifying the names of large datasets as a way of filtering in and out relevant material. For this review, both the reviewers had to use their strong judgment to determine that each review as a whole sufficiently met our study hypothesis and research questions of scoping review. This might have left some gap within the methodological process subject to reviewer bias.

5. Results

5.1. How has "big data" been used in immigration studies?

According to the report "big data in Action for Development", produced by the World Bank Group (López. H, et al. 2014.), there are two basic big data approaches identified in Global context. Among the two approaches, one way of utilizing big data is "for projects or processes which seek to analyze behaviors outside of government or development agencies in order to heighten awareness and inform decision making" (López. H, et al. 2014, Pg.11). The second approach helps in analyzing "behaviors internal to a single institution, such as the government, often to streamline and improve services." (López. H, et al. 2014, Pg.11). Generally speaking, our review finds that most of the data used in published research on immigrants deals with data in the second approach, government level administrative data. There were, however, a handful of studies that utilized non-administrative data to examine certain aspects of the migration experience. The report on the "Big data in Action for Development " (López. H, et al. 2014), gives an apt example on how text analysis of social media data has been able to successfully identify problems of different groups of people such as refugees and also map human migration patterns, utilizing Big Data for developing action projects involving both data scientists and practitioners together for solving different social issues. Topics uncovered in our scoping review on immigration and 'big data' include labour market outcomes, cultural and ethnic diversity, health assessments, and various aspects related to social integration. The following charts and tables reflect some of the major themes and observations we made as we surveyed the existing research.

The majority of immigration studies utilizing big data were published after 2007, but the trend appears to be upward, with more and more published studies using big data than ever before.

Figure 1- Number of Big Data Studies Published by Year of Publication

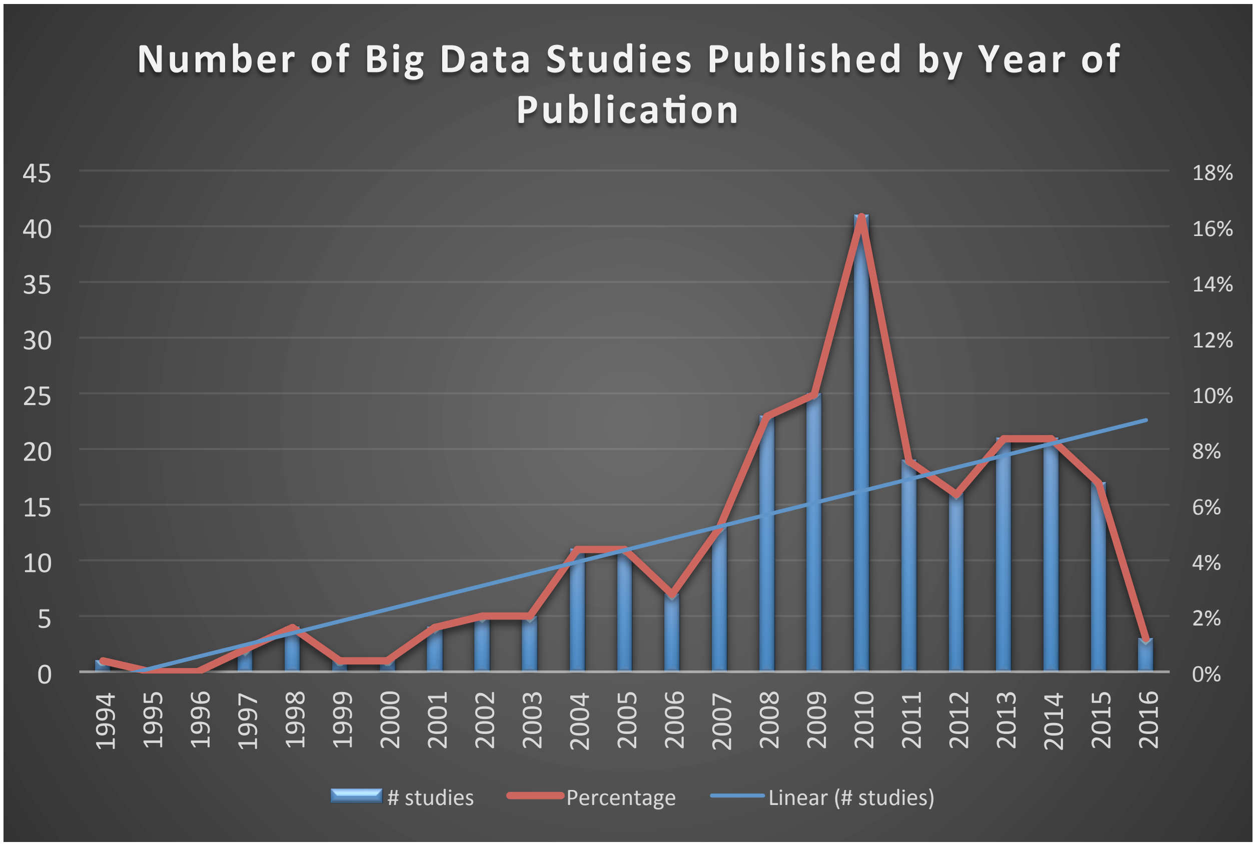
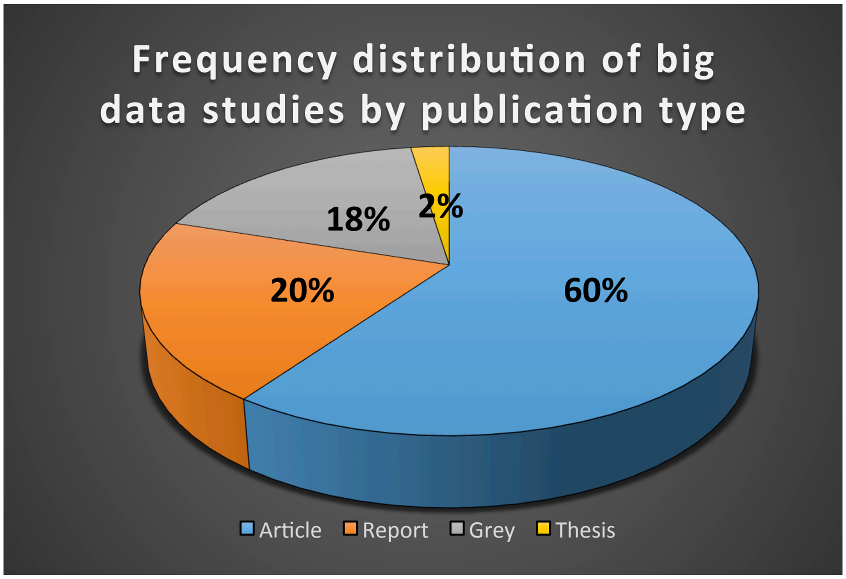


Figure 1 examines the growing trend of publications involving big data and immigration. Approximately 79% (199 out of 251) of the big data research studies identified have been published between 2007 and 2016. In 2010, the largest number of studies involving big data were published, with a total of 41. After 2010, about 20 studies per year involved big data. Clearly, researchers and policy makers are beginning to rely more on big data for their research in the field of immigration studies and the trend seems to be continuing. This is likely a reflection of better access to big data and researchers with better 'big data' skill sets than in previous years.

The majority of immigration research utilizing big data is published in academic journals.

Figure 2- Frequency distribution of big data studies by publication type



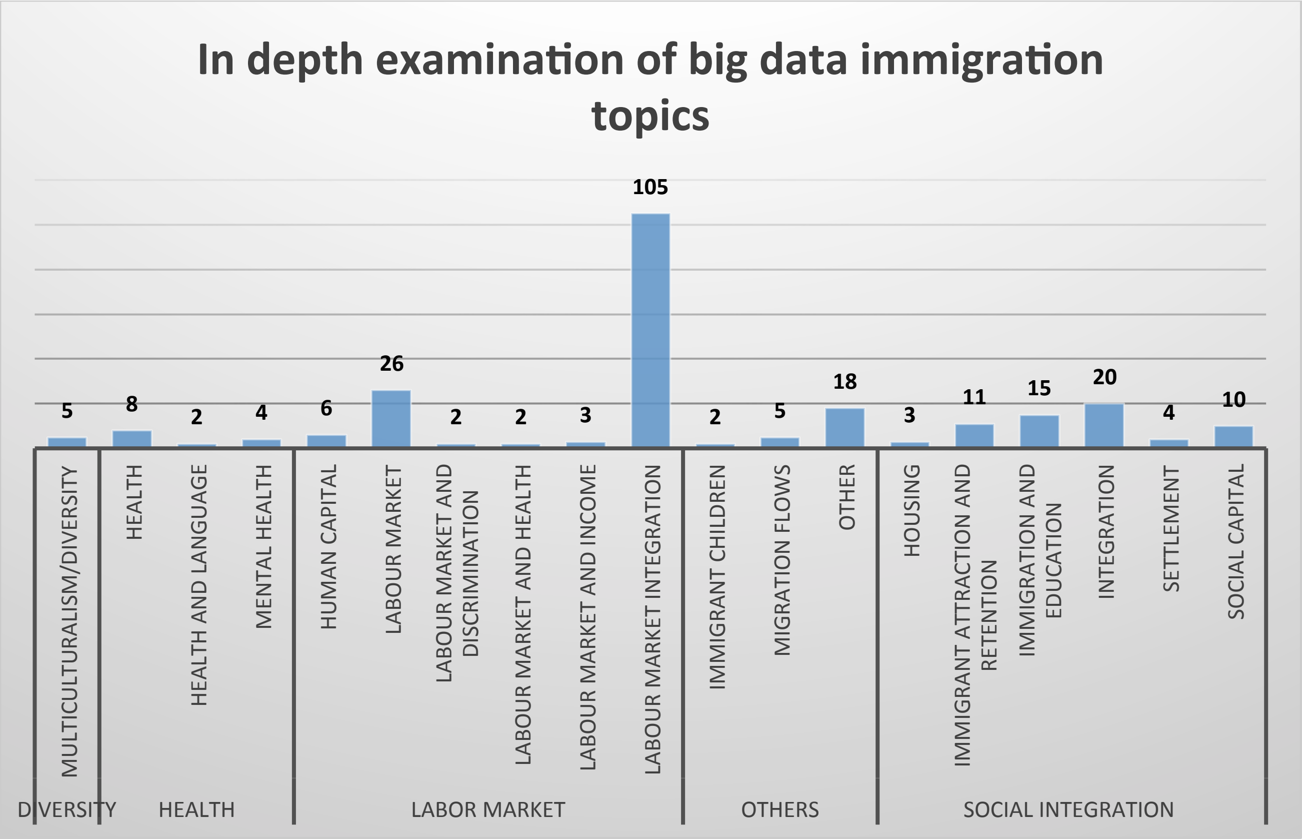
Most of the big data immigration research we could locate has been published as journal articles (60%). Another 20% are reports, grey literature (18%) or doctoral or master's theses (2%) (see Fig.2). Out of the 251 studies, 150 appear in peer-reviewed academic journals. Another 51 studies are government reports and 44 appear in the grey literature (defined as technical and working papers, statistical reports, etc.).

Labour market issues dominate studies using big data.

Within the reviewed literature, the predominant theme was immigrant labour market outcomes with 57% of the reviewed papers (144 out of 251). This is not surprising given that an examination of publication themes of the major peer-reviewed journals in immigration worldwide reveals that labour market studies are the theme of over 60% of all published articles for the past 15 years (Wilkinson, 2015) and is a trend that has continued since the mid-1990s (Li, 1997). The next most popular topic was social integration with 25% (63 out of 251) of the reviewed studies. Health, ethnic and cultural diversity and others make up the remaining 18% of the articles using big data.

Figure 3 presents the major themes and sub-topics of immigration studies that use big data. As mentioned above, the majority of studies utilizing big data in their research have examined the field of immigrant labour market outcomes. Human capital, labour market discrimination, Labour health and jobs, and income are the topics most discussed. Labour market integration is the most predominant sub-theme, not only within the labour market domain, but in all of the research we identified (72.9%). Another 26 studies were categorized broadly as "labour market" and included related topics such as Immigrant Self-Employment and Entrepreneurship, Immigrants and the Dynamics of High-Wage Jobs, Immigrant earnings and economic outcomes. Within the broader theme of "social integration", we find sub-themes such as "housing", "immigrant attraction and retention", "immigration and education", "integration", "settlement" and "social capital" among the most published topics. A 'catch-all' category contains the papers that couldn't be categorized in other ways. It contains research focusing on aspects related to migration and children, the number of immigrant arrivals, and other studies that couldn't be categorized such as a study on "racialized selection of immigrants in Canada" and a study on "Global banking and financial services to immigrants". Within the health theme, physical health was most predominant, with mental health and "the influence of language on various aspects related to health as major themes of research. There were 5 studies which were based on "multiculturalism and diversity" included within the broader category of diversity.

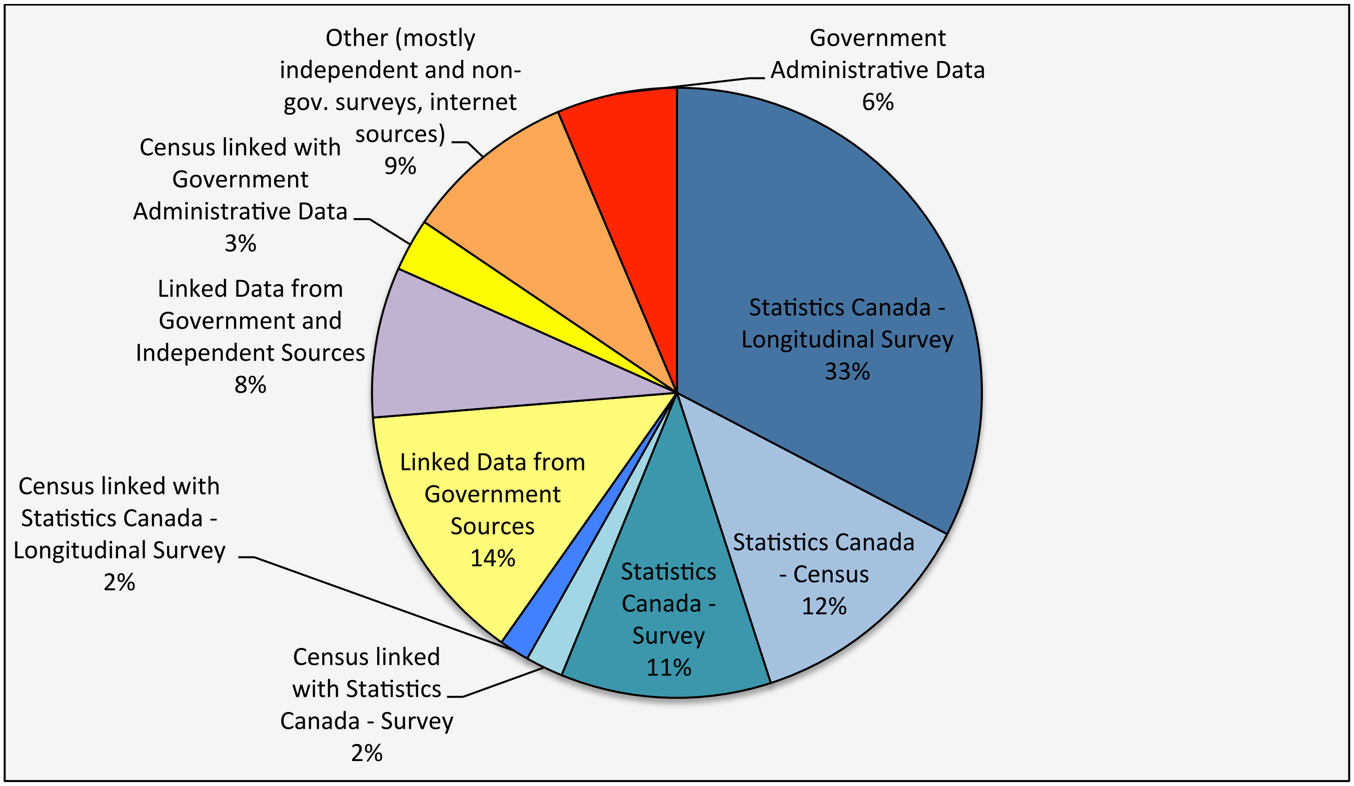
Figure 3- In depth examination of big data immigration topics



Close to 60% of the immigration studies utilizing big data use Statistics Canada surveys as their data source.

In Canada, the federal government funds the bulk of social science research. They are also the main collectors of public use microdata in the forms of surveys. It is not surprising that government data makes up the majority of data sources for immigration and big data research. Almost a third (29%) of all the data sources involved some linked government data, usually between an administrative database and a survey. Another third used longitudinal surveys. Figure 4 also shows that almost 60% of the studies reviewed utilize Statistics Canada data. Only nine percent of the studies use data extracted from independent and non-government surveys and internet sources.

Figure 4 – Types of data sources used in immigration research



The Canadian Census and the Longitudinal Survey of Immigrants to Canada (LSIC) are the datasets most utilized by researchers.

Figure 5 - Number of Data Sources per study

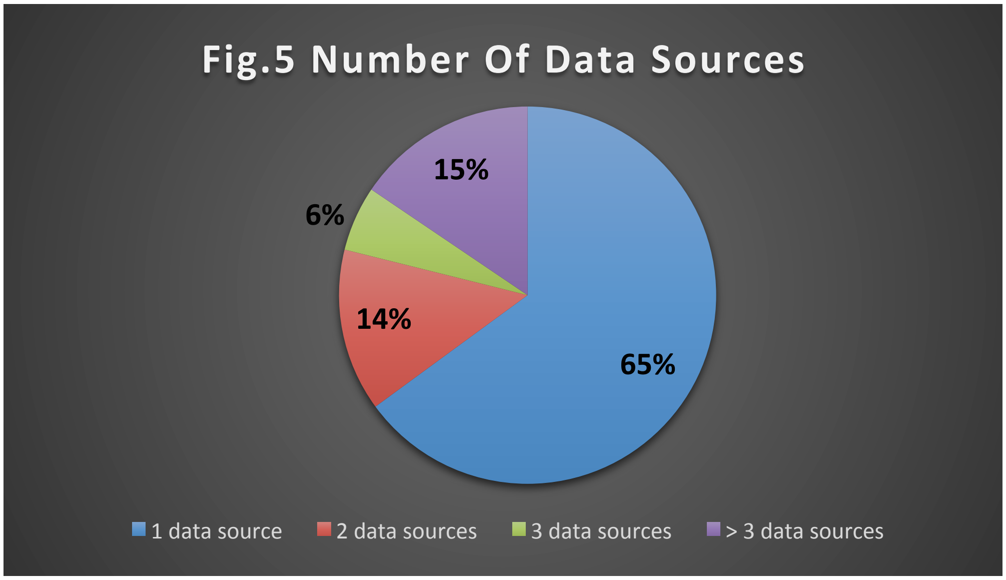


Figure 5 represents the number of big data sources used in the reviewed studies. Among the 251 studies, 163 (65 %) used one data set, 35 (15 %) used two data sets, 14 (6 %) used three data sets while another 39 (15 %) of the studies used more than three datasets. In summary, big data is often not limited to the examination of a single data source, which makes analysis more difficult but more holistic.

Figure 6- Most Popular Datasets used

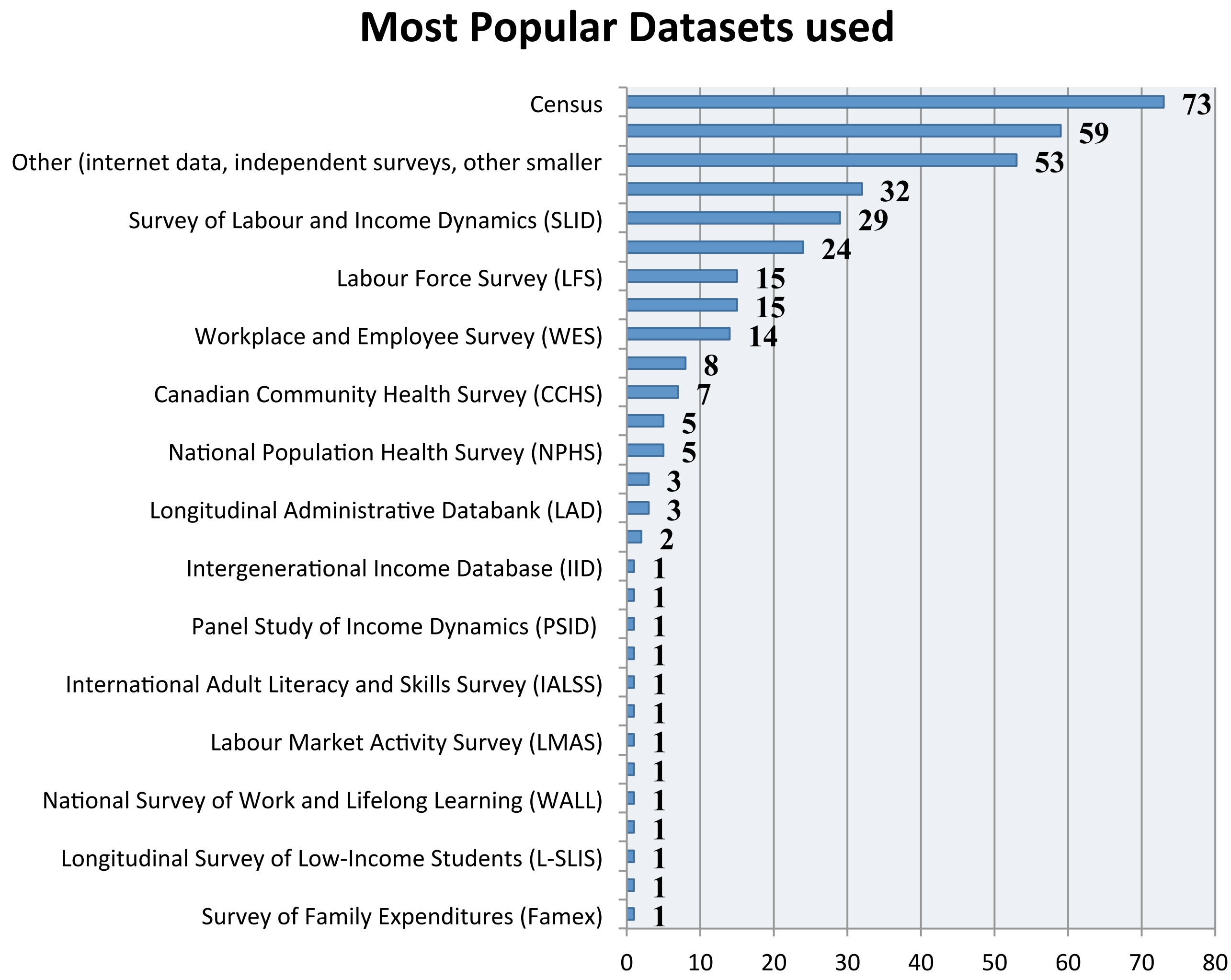


Figure 6 shows the types of databases used by the studies included in our scoping review. Studies using more than one database were counted more than once. The Census data has been used in 73 studies out of 251 (29%) and is by far, the most cited data. This is followed by the Longitudinal Survey of Immigrants to Canada (LSIC), appearing in 24% of the studies captured in the scoping review. This is interesting, given that the survey ended data collection in 2004 but researchers continue to publish from the wealth of this work.

There are databases using internet-based data, independent surveys, data from other countries (for comparative purposes) which were categorized as "others" and were found in fifty three studies Reflecting the preoccupation with labourmarket outcomes, the Survey of Labour and Income Dynamics (SLID)(29 studies), Labour Force Survey (LFS) (15 studies), Workplace and Employee Survey (WES) (14 studies), Financial Security (SFS) (2 studies), and Survey of Family Expenditures (Famex), Youth in Transition Survey (YITS) are also cited as sources of data.

5.2. Challenges and limitations of using big data in immigration studies

Although there are numerous benefits to using big data, we know there are various challenges and limitations and the results of the scoping review revealed many. We have summarized them here and added some of our own ideas.

Experts in big data talk about limitations regarding the sample selection and sampling bias built into big data. Big data sets, despite their large number of participants and variables, have issues of population underrepresentation, especially in administrative data and often contribute to inaccurate and unsettled relationship between selected variables. For example, people who have no contact with the medical system may not appear in administrative databases and skew the results. This has bigger implications for immigration research. Researchers who engage in immigration research face the issue of having to deal with relatively small sample sizes even within large datasets. The National Longitudinal Study of Children and Youth, for example, chronically under sampled immigrant and refugee children for the first few years of its existence. Even LSIC has built-in biases, particularly in the small sample size of adults aged 55 and older (Hochbaum 2013). To be fair, LSIC was intended to follow newcomers at arrival and very few immigrants arrive to Canada who are aged 55 and older. Sampling issues can also occur even in targeted databases. For instance, even though the Longitudinal Survey of Immigrants to Canada (LSIC) contains a large volume of information about the immigrant experience, it is very difficult to examine newcomers from all but the five largest immigrant-sending countries. Even sex differences can be difficult to examine given the small sample size. Pottie and colleagues (2008) describe the recruitment problems with longitudinal studies, particularly with immigrants, as many fear for their future in Canada (they might be deported) and were reluctant to participate in the study. Simone and her research team (2014) complain that LSIC does not include refugee claimants and temporary workers so these groups are largely ignored in research.

Administrative data can be difficult to use, particularly in immigration studies. Often, this data must be linked with the Immigrant Landing File (ILF)—and even though it is a census of all newcomers to Canada, it may not be possible to make positive matches with new data. Individuals who are difficult to match (who may have incorrect identifiers entered into the administrative database, for instance) are not 'matched' and are then excluded from the database. There are known patterns to this type of information—with those arriving as children or as teens more likely to be 'unmatchable' due to identification-based database errors of this sort. Undoubtedly, when ILF is used to link with a Statistics Canada Survey, researchers have a chance to examine data in ways that were not previously possible. This has opened up research on questions related to refugees, who have historically been excluded from analyses or are mixed in with other types of newcomers. While this has been important to advancing our knowledge of this group, there are other problems with linking ILF to surveys meant for the general population. Lightman and her colleagues (2013) had difficulty analyzing the Survey on Labour and Income Dynamics because items pertaining to the immigrant experience, such as country of last job and country of highest level of education, were not asked in that survey. It is frustrating from a Canadian standpoint because nearly one-quarter of the Canadian population was not born here and the failure of general surveys to routinely collect these items embeds even more bias into results. It is a frustration echoed by other researchers investigating other databases (Marzia 2015; Papademetriou et al., 2009).

Another challenge is that some comprehensive big data sets, such as the ones obtained from Statistics Canada, have been reported by users as being nearly impossible to validate the measures and indicators within the data due to missing data, not knowing how to manipulate the information, inconsistency in how the surveys or censuses were collected, difficulty with how to operationalize and measure selected variables (Nikolaos 2002). The impact it has on immigration studies in Canada is that, it influences the kind of studies researchers can do and the kind of questions they can ask.

Storage, management, and sharing of big data remain a challenge for all users of big data. These three issues are connected. In Canada, as elsewhere, many of these large data sets, both administration and survey-based, are located in secure facilities. The protection of data is a big issue in 21st century research and the possibility that the anonymity of participants could be jeopardized if the data were to fall into the wrong hands is real (Franke et al, 2015). With the increased security of data, comes restriction to access, meaning fewer and fewer researchers can access the data. Rigorous security screening is often required before a researcher can access data. They are also constrained by the office hours of the organization of whose data they are using. Students have a particularly difficult time in accessing data, although the Research Data Centres of Statistics Canada have been a golden opportunity for them to access data that only their supervisors could examine in previous decades. One of the unintended consequences of secure access is that researchers outside the government and university find it next to impossible to access data. Without a university-affiliation, most of this data is out of the reach of non-government settlement organizations, groups that can critically analyze this information for practical purposes. It sets up a two-tiered system of access where researchers are increasingly the gatekeepers of data that community organizations could use. Heather and James (2013) point to a related issue which they call “immature technology”. By immature technology, they lament the pace at which big data becomes available to the public as compared to the technology needed to manage and analyze mega-datasets. Alongside immature technology, we would like to add that the number of researchers with the advanced skillsets in both data management and statistics are in short supply so the reality is that even if more big data were to become available, there will be a shortage of qualified people to analyze and interpret the numbers. According to IBM, there will be a shortage of over 4.6 million data and analytical workers in the coming decade.

There are ethical issues with the use of linked data and with administrative data. Often, this data is collected from individuals who are in need of a service such as medical care. No one tells them when they visit their physician (or file their income tax or send their child to school) that the data collected could be used for research purposes in the future. All universities and federally and provincially funded research requires an ethics review prior to the beginning of data collection and many of the data sets used in studies for this scoping review did not seek ethical approval from participants. One notorious example is the IMDB which links the ILF with the tax record file. All immigrants who have arrived to Canada on or after 1980 are part of this dataset.

Data from administrative sources, collected from social media, internet and online communication sources are subject to bias, even if they are presented as a census. Credibility of results is a major issue. Zagheni and his colleagues (2014) point out that information collected from social media can be produced by anyone (or anything such as a cyberbot) without going through the rigorous process of data verification, validation and analysis. There is no valid way for researchers to identify data generated by computer versus human so this sort of data is always suspected.

5.3. Opportunities for immigration research with big data

Despite these problems, there is real potential for big data in advancing future immigration research. In this period of displacement crisis, “accurate and timely statistics are needed to assist migrants effectively” (Marzia, 2015, para. 3). Migration data can be very helpful for settlement organizations to use when developing new programs or expanding on existing ones. One of the biggest challenges that community organizations faced during the arrival of Syrian refugees over a short period of time was not knowing how big their families were. This is where the timely release of big data, much like the handy maps of Syrian arrivals to Canada released by IRCC, can be used to a greater degree. Even basic information such as average family size can help community organizations locate appropriate housing in a more timely fashion. Big data on migrants helps not only in creating “sensible migration policies” but also ensures adequate distribution of resources. If the sharing of these data can be more democratic, with easier access to this vital data, particularly to community organizations, then big data can be harnessed to help design and implement better policies and programs for newcomers in a more timely fashion.

Good research is informed by both policy and practice. Researchers should be encouraged to partner with institutions and organizations to examine questions relevant to the practical needs of the settlement sector. More of this research needs to be done and if these partnerships can be sustained, more data will be used and more articles and reports will be read. Franke et al. (2015) remind us that researchers can benefit from the insight of community settlement service workers, particularly in interpreting the results of data analysis. Since big data is complex, continuously changing and growing, it requires diverse knowledge on data manipulation.

The Syrian crisis has shown, with great clarity, the ubiquitous nature of cell phone use and internet access. Only a few studies have used this data to examine immigration. Although there are inherent risks with using this data, great knowledge can also be gained when we examine it. This reflects how “unprecedentedly large and complex amount of data is being generated in real time every time a call or an online payment is made, or every time people interact on social media, which is usually referred to as big data” (Marzia, 2015, para. 9). Activities using social media, web login history, IP addresses, emails and text messages can be geolocated and turned into migration data using statistical modeling. “Massive datasets and social networking sites provide opportunities to design experiments on a scale that was previously impossible in the social sciences”, (Grimmer, 2015, p.81). When used appropriately, this data can shed light on important aspects of immigration.

In order for immigration research to benefit from the current and future opportunities with a variety of new big data sources, there is a need to train more people in data analysis and statistical modeling. The complex work done by researchers such as Jedwab and Soroka (2014) is one example of how complex these analyses can be. If we invest in additional training of students to handle large datasets, more research can use big data. Renewed training may actually encourage more people to enter the immigration field as more data analysis opportunities become available.

There are some exciting developments concerning linked administrative and survey data relating to immigration studies in Canada. In 2016, IRCC, along with Statistics Canada, announced some new exciting data linkages to be released in the coming 18 months. The ILF will be linked to the following datasets: the Canadian Employer and Employee Dynamic Database (including international students and temporary workers), the 2011 National Household Survey, the 2016 Census of Canada, the Canadian Community Health Survey and the 2013 General Social Survey on Social Identity. IMDB will continue to be released. The new Settlement Outcomes Survey will also be linked to administrative databases. Health records from Ontario, BC and Manitoba will also be linked to the ILF. Some researchers may be able to access iCARE, the administrative database that houses information on all settlement services used by newcomers across Canada (IRCC, 2016). In short, there will be great opportunities in the very near future for researchers and community organizations to learn more about newcomers on a variety of different topics.

6. Knowledge Mobilization

The potential knowledge users of synthesis results include, but are not limited to, researchers in various academic institutions, government officials, policy analysts, and representatives of the integration and settlement sectors, as well as graduate students. The research findings from this report will be disseminated widely to ensure an optimal research uptake from all research knowledge groups. This project utilized three types of knowledge mobilization strategies: integrated, end-of-grant and post-grant. The knowledge mobilization strategies were developed by the research team and were communicated with the advisory panel members.

6.1. Integrated knowledge mobilization.

The Advisory Committee was comprised of three individuals; one representative of the Manitoba provincial government and two representatives of non-governmental organizations (NGOs) in British Columbia and Alberta. The purpose of the Advisory Committee is to assist the research team in the interpretation and the dissemination of the research findings and to provide input and feedback on the reporting. The three individuals represent groups of potential knowledge users and participated in one meeting to provide feedback on the research approach and their input on how this research can benefit their organizations (May). An informal consultation took place via e-mails in August where the preliminary research outcomes were shared with the Advisory Committee inviting the members' feedback. The advisory panel will assist the team in disseminating the results of the research.

6.2. End-of-Grant and post-grant knowledge mobilization.

The research findings will be disseminated in various ways to ensure an optimal research uptake from all research knowledge groups. The following end-of-grant and post-grant knowledge mobilization activities are planned:

- Conduct a meeting with the Advisory Committee (November) to discuss the research findings and the most efficient and practical ways to package and disseminate the research findings to non-academic audiences and front-end big data users.
- Present the research findings at "The Canadian Institute for Identities and Migration (CIIM) & Immigration Research West (IRW) Regional Symposium: Migration and Refuge in Western Canada" on October 21st, 2016.
- Present the research findings at Social Sciences and Humanities Research Council's Fall Forum in Ottawa on November 22, 2016
- Submit a proposal for a presentation at the 19th National Metropolis Conference.
- Submit a publication to the Canadian Diversity, Association for Canadian Studies.

- Produce a web-based resource of the tabulated database used during the study selection and charting the data phases of the research process. All materials produced from this study will be hosted on the Rural Development Institute, Brandon University and Immigration research West, University of Manitoba' websites. Links to the above websites will be distributed through Research Partners Network (e.g. Rural Policy Learning Commons) and Immigration Research West Network.

7. References

Benton, M. (2014). *Smart inclusive cities: How new apps, big data, and collaborative technologies are transforming immigrant integration.* Washington, DC: The Migration Policy Institute.

Papademetriou, D. G., Somerville, W., & Sumption, M. (2009). The social mobility of immigrants and their children. *Migration Policy Institute, Washington*.

Franke, B., Plante, J. F., Roscher, R., Lee, A., Smyth, C., Hatefi, A., ... & Hoffman, M. M. (2015). Statistical Inference, Learning and Models in Big Data. *arXiv preprint arXiv:1509.02900*.

Garcea, J., Jason D., Winston Z., & Wilkinson, L. (2016). Geo-spatial Data of Immigration to Canada's Western and Northern Region. Winnipeg; Immigration Research West. Accessed online at http://gistest.usask.ca/irw/

Grimmer, J. (2015). We are all social scientists now: how big data, machine learning, and causal inference work together. *PS: Political Science & Politics*, *48*(01), 80-83.

Smith, H., A., & McKeen J, D. (2012). Big data and Data Analytics. *CIO Brief*, 18(3), 1-6.

Hilbert, M., & López, P. (2011). The world's technological capacity to store, communicate, and compute information. *science*, *332*(6025), 60-65.

Hochbaum, C. V. (2013). Too Old to Work?: The Influence of Retraining on Employment Status for Older Immigrants to Canada. *Canadian Ethnic Studies*, *44*(3), 97-120.

Huebner, R. A. (2013). A Survey of Educational Data-Mining Research. *Research in higher education journal*, *19*.

Humanitarian Tracker (2016). Syria Tracker. Accessed online at http://www.humanitariantracker.org/syria-tracker

ICT (2016) World Telecommunication/ICT Indicators Database. 12 July 2016. Accessed online at http://www.itu.int/en/ITU-D/Statistics/Pages/stat/default.aspx

International Organization for Migration (2016). Mixed Migration Flows in the Mediterranean and Beyond. Geneva: IOM. Accessed online at http://migration.iom.int/docs/WEEKLY%20Flows%20Compilation%20No26%206%20October%202016.pdf

International Live Statistics (2016) Internet Users by Country. Accessed online at http://www.internetlivestats.com/internet-users/

Immigration Refugees and Citizenship Canada (2016). Map of destination communities and settlement service provider organizations, updated 21 September 2016. Ottawa: IRCC. Accessed online at http://www.cic.gc.ca/english/refugees/welcome/map.asp

Immigration Refugees and Citizenship Canada (2016). Researching Syrian Refugees: Data Availability and Access. Ottawa: IRCC.

Jedwab, J. & Soroka, S. (2014). "Indexing Integration: A Review Of National And International Models". A report prepared for the Department of Citizenship and Immigration Canada by the Association for Canadian Studies (Canadian Institute for Identities and Migration), 30.

Khandor, E., & Koch, A. (2011). *The global city: newcomer health in Toronto*. Toronto Public Health.

Laczko, F. & Rango, M. (2014). Can big data help us achieve a 'migration data revolution'? *Migration Policy Practice* 6(2), 20-29.

Levac, D., Colquhoun, H., & O'Brien, K. K. (2010). Scoping studies: advancing the methodology. *Implementation Science*, *5*(1), 1.

Li, P.S. (1997) A Review of the Academic Literature on Immigration Studies in Canada. Ottawa: Citizenship and Immigration Canada.

López. H, et al. 2014. Big data in Action for Development. The World Bank group. Accessed online at http://live.worldbank.org/sites/default/files/Big%20Data%20for%20Development%20Report_final%20version.pdf

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). Big data: The next frontier for innovation, competition, and productivity. Accessed online at http://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation

McNulty, E. (2014). "Understanding big data: the 7 vs" Dataconomy. Accessed online at http://dataconomy.com/seven-vs-big-data/

Mendez, P., Wyly, E., & Hiebert, D. (2011). Landing at home: Insights on immigration and metropolitan housing markets from the longitudinal survey of immigrants to Canada.

Liodakis, N., & Satzewich, Victor. (2002). *The Vertical Mosaic Within: Class, Gender and Nativity within Ethnicity,* ProQuest Dissertations and Theses.

OECD (2013). Exploring Data-driven Innovation as a New Source of Growth: Mapping the Policy Issues Raised by 'big data'. Accessed online at http://www.oecd-ilibrary.org/science-and-technology/exploring-data-driven-innovation-as-a-new-source-of-growth_5k47zw3fcp43-en

Pandey, M., & Townsend, J. (2011). *Provincial nominee programs: An evaluation of the earnings and retention rates of nominees*. Prairie Metropolis Centre.

Pottie, K., Ng, E., Spitzer, D., Mohammed, A., & Glazier, R. (2008). Language proficiency, gender and self-reported health: an analysis of the first two waves of the longitudinal survey of immigrants to Canada. *Canadian Journal of Public Health/Revue Canadienne de Sante'e Publique*, 505-510.

Rango, M. 2015. "How big data Can Help Migrants". World Economic Forum. Extracted from online source https://www.weforum.org/agenda/2015/10/how-big-data-can-help-migrants/

Rathnam, L. (2015) "Can big data help the Syrian conflict?" icrunchdata.com October 8. Accessed online at https://icurnchdata.com/big-data-help-syrian-conflict/

Ratti, C. & Helbing, D. (2016). "The hidden danger of big data" *The Straits Times*, September 12, Accessed online at www.straitstimes.com/opinion/the-hidden-danger-of-big-data.com

Simone, D., & Newbold, K. B. (2014). Housing trajectories across the urban hierarchy: analysis of the longitudinal survey of immigrants to Canada, 2001–2005. *Housing Studies*, *29*(8), 1096-1116.

Sweetman, A., & Warman, C. (2012). The structure of Canada's immigration system and Canadian labour market outcomes. *Queen's Economics Department Working Paper No*, *1292*.

UN Global Pulse. 2014. Estimating Migration Flows Using Online Search Data, Global Pulse Project Series no. 4.

Van Rijmenam, M. (2014). Why the 3Vs are not sufficient to describe big data. Datafloq. Accessed online at http://dataconomy.com/seven-vs-big-data/

Wilkinson, L. (2015) "Immigration research and policy developments: what do we know and where are we headed?" plenary panel paper presented at the Canadian Sociological Association Annual Conference, Ottawa, University of Ottawa

Zagheni, E., Garimella, V. R. K., & Weber, I. (2014). Inferring international and internal migration patterns from twitter data. In *Proceedings of the 23rd International Conference on World Wide Web* (pp. 439-444). ACM.