Review Article

iMedPub Journals
http://www.imedpub.com/

Health Science Journal
ISSN 1791-809X

2017
Vol.11 No.3:510

DOI: 10.21767/1791-809X.1000510

Methodological Challenges to Collecting Valid Research Data: Researching English Language Barrier Populations in Canada

Emmanuel Ngwakongnwi

Research and Graduate studies, School of Nursing, University of Calgary in Qatar, Qatar

Corresponding author: Emmanuel Ngwakongnwi, Assistant Professor and Research Coordinator, Research and Graduate Studies, School of Nursing, University of Calgary in Qatar, P. O. Box 23133 Doha, Qatar, Tel: +974 4406 5239; Fax: +974 4406 5299; E-mail: engwakon@ucalgary.ca

Received date: 01 June 2017; **Accepted date:** 13 June 2017; **Published date:** 20 June 2017

Citation: Ngwakongnwi E. Methodological Challenges to Collecting Valid Research Data: Researching English Language Barrier Populations in Canada. Health Sci J 2017, 11: 3.

Abstract

Background: Canada has two official languages, English and French. French is the main language in Quebec while English is widely spoken in other provinces and territories. Canadian census data has shown that there is a significant portion of the population in the English speaking provinces that identify French as their mother tongue; and prefer to be served in French when they seek healthcare. This means that they are not able to speak, read, write, or understand the English language in a way that enables them to interact with others daily in English, let alone participate in research conducted in English. Thus the designation, English Language Barrier (ELB). The ELB populations can be difficult to sample for research purposes especially when language affiliation becomes an important identifying variable, thus, can be described as a potentially hard to-reach populations. Inadequate sampling compromises quality of data from which inferences are derived.

Aim: The purpose of this paper is to present challenges to collecting valid research data and suggest ways by which such challenges can be overcome in ELB populations in general.

Methods: To achieve this, I start by defining ELB populations, followed by a brief description of what constitutes valid data. Subsequently, I discuss evidence on data collection practices and identify major challenges. In the discussion, I present ways by which these challenges can be mitigated; and conclude with a summary of the evidence, the challenges, and suggestions to improve data collection in general.

Conclusions: Collection of valid research data among ELB populations can be enhanced by using innovative sampling techniques such as respondent driven sampling, reducing bias and improving on the questionnaire.

Keywords: Sampling techniques; Health care; Health services

Introduction

Data constitutes a vital aspect of research. Given the high costs and time needed to collect data from the entire population, researchers often sample a proportion of the population for study. Usually, the research question determines the method to be used. Valid data, collected from trustworthy sources, using appropriate methods is needed to make inferences and to ensure that findings can be generalized to the population. Unfortunately, data collection is not entirely a smooth process. Upon securing funding, researchers have to get their studies approved by research ethics boards; execution requires rigorous planning, often involving many stakeholders; data sources have to be identified, and may include existing data (records, registers). Where existing data is lacking, new recruits (human subjects) are often solicited to participate voluntarily in studies. The various stages of the research process present certain challenges that may impact the collection of valid data for research on human subjects.

The purpose of this paper is to present challenges to collecting valid research data and suggest ways by which such challenges can be overcome in English Language Barrier (ELB) populations in Canada. To achieve this, I start by defining ELB populations, followed by a brief description of what constitutes valid data. Subsequently, I discuss evidence on data collection practices and identify major challenges. In the discussion, I present ways by which these challenges can be mitigated; and conclude with a summary of the evidence, the challenges, and suggestions to improve data collection in general.

Who are English language barrier populations?

In 2011, 20.6% (6.8 million people) reported a mother tongue other than English or French, and 6.2% of these Canadians spoke a language other than English or French as their sole home language [1]. This, and French only speakers

(Francophones), constitutes what I refer to as 4 English language barrier (ELB) populations. This means that they are not able to speak, read, write, or understand the English language in a way that enables them to interact with others daily in English, let alone participate in research conducted in English. It is worth mentioning that Canada places much emphasis on official languages, designating Official Language Minority communities (OLMC) as Anglophones living in Quebec and Francophones living in provinces and territories outside of Quebec [2]. ELB populations can be difficult to sample for research purposes especially when language affiliation becomes an important identifying variable, thus, can be described as a potentially hard-to-reach populations. Inadequate sampling compromises quality of data from which inferences are derived.

Indices of valid data

In the conduct of epidemiologic and health services research, measurement error is potentially a major problem that may invalidate the results of otherwise well-designed studies. Concepts used to evaluate the quality of measurements include validity and reliability [3]. Validity refers to the extent to which the measurement represents the true value of the attribute being assessed. Assessment of validity is achieved through calculation of quantitative indices of the accuracy of measurement. For discrete variables, two aspects of the accuracy of measurement include: 1) sensitivity (the proportion of those who truly have the characteristic that are correctly classified as having it by the measurement technique); and 2) specificity (the proportion of those who truly do not have the characteristic that are correctly classified as not having it by the measurement technique). Reliability refers to the extent to which results of a measurement can be replicated. Additional forms of reliability- inter-rater reliability (e.g. for assessing agreement in diagnosis) or test-retest reliability (e.g. for comparing measures of blood pressure) can be distinguished. The Kappa coefficient and the correlation coefficient of reproducibility are some of the indices used for the quantification of reliability [3].

Evidence basis for current data collection practices

Health data refers to unprocessed numbers or observations [4]. When analyzed, they become health information; and can be interpreted to inform health policy decision making. According to Spasoff 4 "we appear to have many health data, considerably less health information...partly because it is often easier to collect more data than to analyze and interpret those that are already available". Thus data collection is a recurring process as old data may no longer be relevant or politically convincing. The following examples support the need for recurrent data collection initiatives in ELB populations:

Issues of definition: The definition of language barrier is based on language–which renders difficulty with generating a population representative sample. For example, in the case of francophones, determining a Francophone population requires having information on mother tongue, place of birth, residence and/or language-knowledge of French. Telephone directories contain names of individuals but lack the information required to define populations. Geocoding and surname analysis are alternative methods for defining populations. Geocoding involves using addresses of individuals to identify small areas where they live and linking this information to other databases [5] e.g. Canada Census data to infer their likely ethnicity based on ethnic composition in the area. The accuracy of geocoded estimates of ethnicity largely depends on the extent of racial and ethnic segregation in the geographic areas considered [5-7]. Higher proportions of minority groups living in racially segregated areas yield higher positive predictive value of geocoded estimates meanwhile lower proportions yield higher false positive value.

Surname analysis uses an individual's last name to estimate the likelihood that the individual belongs to a particular ethnic group [7,8]. The prevalence of members of a particular ethnic group in the community has a powerful effect on surname accuracy. In the case of francophones, both geocoding and surname methods may not be reliable and valid for their identification; especially if they are integrated in society and their residences are disaggregated. Moreover, surname reflects ancestry origin but does not necessarily reflect capability of speaking French; high rates of intermarriage between Francophones and other ethnic groups may introduce error.

Non-available gathered data: In Canada, information about immigrants as well as Canadians is collected and securely stored by various government agencies. However, such information is not routinely made available to researchers due to regulatory privacy issues. Administrative hospital data may offer some solutions since it is routinely collected and stored at provincial level. Unfortunately, neither ethnicity nor language variable that can be used to define population subgroups is collected in these administrative data sources [9]. This limits their use for studies involving ELB populations. In a study in Calgary, the participants endorsed collection of ethnicity data in hospitals [9]. Yet, the findings have not been put in practice.

Survey methods need validation to ensure valid data: Depending on the hypothesis being tested and the method employed, tools for data collection need validation prior to being used with ELB populations. For example, measures of health and health status (perceived Health With likert- Type Scale, EQ-5D, and Numbers of Chronic conditions) have been shown to be less applicable to Chinese compared to white Canadians [10]. This means that data collection tools initially developed and used in other populations have to be validated in ELB populations in order that results obtained are valid.

Population distribution patterns: Clusters of populations are sometimes sampled for research. This is meaningful when the characteristics of the people living in the cluster are the same. This may not be the case for ELB populations if they are widely distributed. This is clearly illustrated in Francophone Canadians living in a minority situation. Unlike in Quebec where Francophones have been easily sampled for health studies [11] because of their majority status in community dwellings, it is nearly impossible to draw a random sample of Francophones outside Quebec because in some communities, they are sparsely distributed [11].

Challenges to collecting valid research data

Sampling difficulties: Sampling is central for recruiting participants to epidemiologic studies. The goal of sampling is to optimize representativeness of the sample. The challenge posed by lack of a sampling frame sometimes lead researchers to analyze a few cases [12] or survey in a convenience sample [13,14]. Although such study designs provide some information, the study results are suspect because of the likely biased sample, which leads to biased results (internal validity) and lack generalizability (external validity) to the target population. Non probability sampling techniques, such as convenience sampling have been widely used in the last two decades; however, their popularity (in epidemiologic research) has dwindled because of the inherent biases. Though described as sampling techniques, they should not be confused with the process of sampling in epidemiologic research as they are merely recruitment strategies.

Examples of non-probability sampling techniques include: Venue-based time-space sampling (also called time-location or venue-day-time) is used to recruit study participants at a location such as church, conference or school [15,16]. This type of sampling may allow access to a large number of people with relatively low cost. However only population members who are readily accessible are found and those who are not regularly attending these locals are missed. The recruited participants have common features related to the event or venue, such as religious beliefs and may be significantly different from non-participants of the event in factors of interest.

Snowball sampling: Through this method, participants are recruited by asking individuals to refer those they know, and these individuals, in turn, refer those they know and so on [15]. The sampling continues until the target sample size is obtained. This method has been widely utilized. However, because the respondents are not randomly selected, must be typically found by survey staff, and are dependent on the subjective choices of the first respondents, snowball samples are less likely to provide the basis for valid generalizations to the populations from which the sample was drawn [15]. Thus use of such recruitment strategies in studies of ELB populations may lead to invalid data.

Bias in research: A discussion on collecting valid research data maybe incomplete if the impact that various biases have on results is omitted. A study is considered valid only when three alternative explanations (bias, confounding, random error) have been eliminated [17]. If these alternative explanations are out, the investigators may conclude that the measure is true, and that the study has internal validity. Two main forms of bias exist in epidemiologic research: Selection bias (an error due to systematic differences in characteristics between those who take part in a study and those who do not); and information or observation bias (a flaw that arises from systematic differences in the way information on measures ‘exposure and disease’ is obtained from study groups) [17].

Given that most epidemiologic and health services data is collected by way of survey, biases inherent to this data collection approach constitute a challenge to data quality. Dillman [18] has identified four potential sources of error in mail surveys to include sampling error, noncoverage error, nonresponse error, measurement error. Mail surveys have been widely used in research. However, concerns over response rates (proportion of responders to eligible non responders) have led to increasing use of telephone or mixed mode surveys [19]. The biases involved in use of such methods include response and nonresponse bias [20,21]. 1) Response bias- bias in the ways in which the questions themselves are answered. The respondents may answer in socially desirable ways, repeatedly endorse items regardless of content, expend little effort in the interpretation and answering of questions, avoid extreme response options, or exaggerate in their answers [22]. 2) Nonresponse bias-refers to the bias that exists when respondents to a survey are different from those who did not respond in terms of demographic or attitudinal variables. Nonresponse bias can take two forms-total non-response refers to individuals failing to return the survey at all (or participate in telephone survey), while unit or item non response indicates that the survey was returned incomplete [22]. Unit or item nonresponse introduces missing values in survey data. Use of mixed modes in a single survey (e.g. mail followed by telephone) is in itself biasing, as different modes may impact responses [18].

Inappropriate data collection tools: Most epidemiologic studies use questionnaires to obtain at least some data from participants regarding their exposure to possible risk factors, disease occurrence, and confounding variables. A questionnaire “is a written document used to obtain information from respondents, regardless of whether it is self-administered or administered by an interviewer” [3]. Questionnaires pose a lot of challenges to obtaining valid research data.

These challenges may come from the design of the questionnaire, or may relate to the individuals themselves. For example, data obtained from questionnaires present special problems of measurement in case-control studies because information on exposure must often be obtained only after the disease has manifested itself, sometimes decades after the relevant exposure has taken place [3]. The imperfect memory of individuals about exposures will diminish the quality of data obtained.

Overcoming challenges to collecting valid data

This paper has presented rational for current data collection practices for studies involving ELB in Canada. This section of the paper presents strategies to mitigate and improve on the challenges.

Use innovative sampling techniques: When deciding on sampling methods, it is important to keep in mind that while we strive to save time and money, the choice should be of methods that give the highest degree of accuracy and precision for a given amount of money. Where a sampling frame exists, probability sampling needs to be employed to ensure random selection of participants to a study. Numerous probability sampling techniques (simple random, systematic, stratified, cluster, multistage and ratio sampling) are described in detail elsewhere [3]. An innovative recruitment strategy is through Respondent Driven Sampling (RDS) [23]. RDS is a recent development in sampling methodology based on snowball methods. In RDS, participants are recruited by asking individuals to refer those they know, and these individuals, in turn, refer those they know and so on. RDS was started to overcome some of the limitations of convenience sampling by combining snowball sampling with a mathematical system for weighting the sample to compensate for its not having been drawn as a simple random sample [24]. This technique can be tested on ELB populations and compared with a 'gold standard' technique to assess its performance.

Reducing bias/random error: Random error results from measurement error and sampling variability. Although measurement error can seldom if ever be eliminated, an appreciation of methods for minimizing their impact on results can contribute greatly to the quality of epidemiologic studies and to the appropriateness of the conclusions drawn from them. Most of the challenges posed by biases can be avoided during study design [3]. For example, selection bias can be avoided by using the same criteria for selecting cases and controls, obtaining high participation rates, and taking diagnostic and referral practices into account when designing a study [3,17]. Observation bias can be avoided by masking interviewers and subjects to the study hypothesis (interviewer and recall bias); using a control group that is composed of diseased individuals (recall bias); carefully designing the study questionnaire (interviewer and recall bias); relying on non-interview data (interviewer and recall bias); using multiple measurements, the most accurate information source, and sensitive and specific criteria to define exposure and disease (misclassification) [17]. Though not explored in detail in this context, confounding (mixing of effects between the exposure, the disease, and a third variable that is termed a confounder) can be controlled for at the design stage by way of randomization, restriction, matching [17].

Given that sampling error results from heterogeneity on the survey measures among members of the population, this error can be reduced by increasing sample size [21]. In mail surveys, quadrupling sample size decrease sampling error by one half [18]. Efforts to reduce nonresponse error in mail surveys have focus on improving response rate [18,22], the generally accepted indicator of nonresponse error. Examples of tested procedures to improve response rates include: financial incentives, material incentives, follow-up reminders, timing of follow-ups, personalization of correspondence, anonymity of response, questionnaire layout, questionnaire length, color of questionnaire, type of outgoing postage, content of cover letter, source of survey sponsorship [3,18,19]. Ngo-Metzger et al. [25] showed that use of mixed modes including telephone interviews and mail surveys with phone reminder calls improved response rates in LEP Chinese and Vietnamese Americans. While high response rates are appreciable, it is worth noting that a low response rate does not necessarily entail nonresponse error, as those who respond to a survey may not differ in any measurable way from those who do not respond [18].

Improving the questionnaire: Measurement error results from the inability of individuals to provide accurate information or a motivation for whatever reason to provide inaccurate information. It may also result from characteristics of the question (e.g. a question phrased so that it cannot be answered correctly or of the questionnaire, (e.g. the order in which questions are presented). Using more than one mode may not be advisable if the results from the two modes cannot be equated because of differences in who responds (nonresponse bias) and in how they respond (response bias). This can induce measurement error [26].

I have focused my discussion on methodological challenges to collecting valid research data. Other practical issues such as incorporating the language spoken by ELB populations while designing studies, and use of community leaders [27], will encourage participation and in turn improve the quality of data. In survey research, the questionnaire is central to collecting quality data and should be well designed. Details of common problems with questionnaires and how to improve them are discussed elsewhere [3].

Conclusion

Collecting valid research data is vital for making inferences on ELB. The challenges encountered in collecting valid research data can be categorized into:

1) **Issues of definition:** The definition of language barrier is based on language –which renders difficulty with generating a population representative sample. This information is not often collected. Geocoding and surname analysis methods are often used to define populations; however both methods may not be reliable and valid for identifying ELB populations.

2) **Non-available gathered data:** Information collected nationally on Canadians is not made available to researchers due to regulatory privacy issues. Administrative hospital data may offer some solutions but, neither ethnicity nor language variable that can be used to define population subgroups is collected.

3) **Survey methods need validation to ensure valid data:** For example, measures of health and health status have been shown to be less comparable across ethnicities. Surveys have been used frequently to frequent use of surveys to collect data for research purposes. However, sampling difficulties resulting from lack of a sampling frame, use of non-probabilistic sampling techniques; biases and random error; and use of inappropriate data collection tools, pose a challenge to collecting valid data.

The following strategies will improve the quality of research data: i) Use innovative sampling techniques- For example respondent driven sampling is a recruitment technique which represents a methodological advance over traditional snowball-sampling because interviewees are given incentives, and recruit the same number of individuals, and mathematical weights are generated to improve representation of the population. Where a sampling frame is available, probability sampling needs to be employed to ensure random selection of participants to a study. ii) Reducing bias- Most of the challenges posed by biases can be avoided during study design. For example, selection bias can be avoided by using the same criteria for selecting cases and controls, obtaining high participation rates, and taking diagnostic and referral practices into account when designing a study. iii) Improving the questionnaire: Improving the way questions are phrased, the format and order of questions among other things will reduce measurement error. Obtaining valid data is vital for making inferences in health services and epidemiologic research. Addressing the methodological issues, and practical issues such as incorporating the language spoken by English language barrier populations while designing studies, will improve the quality of data.

Conflict of Interest

The author has no conflicts of interest.

References

1. http://www12.statcan.gc.ca/census-recensement/2011/as-sa/98-314-x/98-314-x2011001-eng.cfm
2. http://laws-lois.justice.gc.ca/eng/acts/O-3.01/
3. Kelsey J, Whittemore A, Evans A, Thompson W (1996) Methods in observational epidemiology. Oxford University Press, Oxford.
4. Spasoff RA (1999) Epidemiologic measures for health policy.
5. Fremont AM, Bierman A, Wickstrom SL, Bird CE, Shah M, et al. (2005) Use of geocoding in managed care settings to identify quality disparities. Health Aff 24: 516-526.
6. Kwok RK, Yankaskas BC (2001) The use of census data for determining race and education as ses indicators: A validation study. Ann Epidemiol 11: 171-177.
7. Fiscella KFM (2006) Use of geocoding and surname analysis to estimate race and ethnicity. Health Research and Ed Trust 41: 1482-1500.
8. Quan H, Ghali WA, Dean S (2004) Validity of using surname to define Chinese ethnicity. Can J Public Health 95: 314.
9. Quan H, Wong A, Johnson D, Ghali WA (2006) The public endorses collection of ethnicity information in hospital: implications for routine data capture in Canadian health systems. Health Policy 1: 55-64.
10. Leung B, Luo N, So L, Quan H (2007) Comparing three measures of health status perceived health with Likert-Type scale, EQ-D, and number of chronic conditions in Chinese and white Canadians. Med care 45: 610-617.
11. Zunzunegui MV, Konq A, Johri M, Beland F, Wolfson C, et al. (2004) Social networks and self-rated health in two French-speaking Canadian community dwelling populations over 65. Social Sci Med 58: 2069-2081.
12. Villalon L, Leclair CA (2008) A participatory approach for the prevention of type 2 diabetes for francophone youth of New Brunswick. Can J Diet Pract Res 65: 15-21.
13. DeWit DJ, Beneteau B (1999) Predictors of the prevalence of tobacco use among francophones and anglophones in the province of Ontario. Health Educ Res 14: 209-223.
14. Marmen L, Delisle S (2003) Healthcare in French [11-008].
15. Magnani R (2005) Review sampling hard-to-reach and hidden populations for HIV surveillance. AIDS 19: S67-S72.
16. Muhib BF (2001) A venue-based method for sampling hard-to-reach populations. Public Health Reports.
17. Aschengrau A (2008) Seage essentials of epidemiology in public health.
18. Dillman DA (1992) The design and administration of mail survey. Annu Rev Sociol.
19. Krosnick JA (1999) Survey research. Annu Rev Psychol 50: 537-567.
20. Caslyn RJ, Winter JP (2003) Understanding and controlling response bias in need assessment studies. In: Linda SJ, Gilmartin SK, Bryant AN (eds.). Assessing response rates and nonresponse bias in web and paper surveys. Research in higher education 44: 409-430.
21. Johnson LC, Beaton LC, Murphy S, Pike K (2000) Sampling bias and other methodological threats to the validity of health survey research. Int J Stress Management 7: 247-267.
22. Linda SJ, Gilmartin SK, Bryant AN (2003) Assessing response rates and nonresponse bias in web and paper surveys. Research in Higher education 44: 409-430.
23. Heckathorn D (1997) A new approach to the study of hidden populations. Social Problems 2: 174-199.
24. Heckathorn D (2002) Respondent -Driven Sampling II: Deriving valid population estimates from chain referral samples of hidden populations. Social Problems 49: 11-34.
25. Ngo-Metzger N, Kaplan SH, Sorkin DH, Clarridge BR, Phillips RS (2004) Surveying minorities with limited English proficiency. Does data collection method affect data quality among Asian American? Medical Care 42: 893-900.
26. Dillman DA, Tarnai J (1991) Mode effects of a cognitively designed recall question: A comparison of answers to telephone and mail surveys. In: Dillman DA (ed.). The design and administration of mail survey. Annu Rev Sociol 1992: 21.
27. Keyzer JF, Melnikow J, Kuppermann M, Birch S, Kuenneth C, et al. (2005) Recruitment strategies for minority participation: Challenges and cost lessons from the power interview. Ethnicity and Disease 15: 395-406.