Estimating patient demographic profiles from practice location

Michael Shortt MA William Hogg MSc MCISc MD FCFP Rose Anne Devlin PhD
Grant Russell MBBS FRACGP MFM PhD Laura Muldoon MD MPH FCFP

Abstract

Objective To test the accuracy of imputing a practice population's average socioeconomic characteristics (such as average education levels and average income) using census data centred on the location of the practice.

Design Comparison of census data with survey data collected in primary care offices.

Setting Ontario.

Participants A cross-sectional sample of patients from 116 urban practices.

Main outcome measures Patient data were compared with census data at different levels of aggregation using mean absolute relative error (ARE), median ARE, and Spearman rank correlations.

Results A total of 4413 patient surveys were collected. Differences between patient profiles and census data were large. Most mean AREs were clustered between 0.70 and 0.80, and median AREs were as high as 1.67. Correlations were low ($\rho = 0.02$) to moderate ($\rho = 0.48$). These results held across both levels of aggregation.

Conclusion The use of imputation techniques based on practice location is inadvisable, given the large differences that were observed.

EDITOR'S KEY POINTS

- Socioeconomic status data on patients are often missing from common health data sources, requiring researchers to gather new data or use the socioeconomic status characteristics of patients' residential neighbourhoods.

- A potentially attractive data gathering strategy is to estimate practice demographic profiles (average age of patients, proportion of male patients, etc) by using census data centred on the location of the practice. However, the present study found important differences between census data and data generated by practice-level surveys. These differences were large enough to render practice-centred imputation invalid.

- A practice's community, if it can be defined at all, is unlikely to be a geographically bounded entity; most patients live outside the geographic neighbourhoods of their primary care practices.

This article has been peer reviewed.
Can Fam Physician 2012;58:414-9

Estimer le profil démographique des patients à partir du lieu de pratique

Michael Shortt MA William Hogg MSc MCISc MD FCFP Rose Anne Devlin PhD
Grant Russell MBBS FRACGP MFM PhD Laura Muldoon MD MPH FCFP

Résumé

Objectif Vérifier avec quelle précision on peut déterminer les caractéristiques socioéconomiques moyennes d'une clientèle (comme les niveaux moyens de scolarisation et le revenu moyen) à partir des données d'un recensement portant principalement sur le lieu de pratique.

Type d'étude Comparaison des données du recensement avec celles d'une enquête effectuée dans des cliniques de soins primaires.

Contexte L'Ontario.

Participants Un échantillon transversal des patients de 116 établissements de pratique urbains.

Principaux paramètres à l'étude On a comparé les données des patients à différents niveaux d'agrégation en utilisant la moyenne de l'erreur relative absolue (ERA), la médiane de l'ERA et des corrélations de Spearman.

Résultats Un total de 4413 enquêtes sur des patients ont été recueillies. Il y avait de grandes différences entre les profils des patients et les données du recensement. La plupart des ERA moyennes étaient regroupées entre 0,70 et 0,80 et les ERA médianes étaient aussi élevées que 1,67. Les corrélations étaient de faibles ($\rho = 0{,}02$) à modérées ($\rho = 0{,}48$). Ces résultats étaient valables aux 2 niveaux d'agrégation.

Conclusion Compte tenu des grandes différences observées, il n'est pas conseillé d'utiliser des techniques d'attribution basées sur le lieu de pratique.

POINTS DE REPÈRE DU RÉDACTEUR

- Les données sur le statut socioéconomique des patients sont souvent absentes des sources habituelles de données sur la santé, ce qui oblige les chercheurs à recueillir de nouvelles données ou à utiliser les caractéristiques du statut démographique du voisinage résidentiel des patients.

- Une façon potentiellement intéressante de recueillir des données consiste à estimer les profils démographiques des clientèles (âge moyen des patients, proportion de patients masculins, etc.) en se servant de données de recensement centré sur le lieu de pratique. Toutefois, cette étude a trouvé d'importantes différences entre les données du recensement et celles obtenues des enquêtes effectuées au niveau de l'établissement de pratique. Ces différences étaient suffisamment importantes pour rendre invalides les attributions basées sur le lieu de pratique.

- Une communauté de pratique, pour autant qu'on puisse la définir, a peu de chances d'être une entité possédant des limites géographiques; la plupart des patients habitent en dehors du voisinage géographique de leur établissement de soins primaires.

Cet article a fait l'objet d'une révision par des pairs.
Can Fam Physician 2012;58:414-9

An important trend in health policy research has been an increasing focus on the role of the social determinants of health.[1] Unfortunately, theoretical progress in this area has rapidly outpaced data collection efforts. As a result, many variables of interest to researchers—ranging from population density to household income to education levels—are unavailable in common data sources such as electronic medical records or health administrative databases. This increases the difficulty of studying the social determinants of health because data on the outcome under investigation (health) are typically separated from data on its causes (social determinants). In the context of primary care research, a project examining the incidence of type 2 diabetes would likely have access to information on biomedical risk factors such as age or obesity, but not to information on known socioeconomic risk factors such as income levels,[2] ethnic origin,[3] or education.[4]

There are 3 solutions to this fragmentation of the field's data sources. One option is to collect new information directly from patients to acquire a broader array of social, economic, and demographic variables. This approach is expensive and time-consuming, and it is not always possible to integrate the resulting information with existing databases. Alternatively, patient characteristics can be estimated using census data centred on the patient's place of residence (personal income might be imputed with average neighbourhood income, for example). Where residence data (typically postal codes) are available, this method has proven to be both popular and effective.[5-8]

In studies that use practices as the unit of analysis, a third option exists. Practice demographic profiles (average patient income, average patient education levels, proportion of patients aged older than 65 years) can be estimated by using census data centred on the location of the practice.[9] This approach is convenient and inexpensive because postal codes for practices are readily available; hence, practice-centred census data are substantially easier to collect than patient-centred census data.

This study tested the accuracy of imputing practice-wide demographic averages using the demographic data centred on the practice's neighbourhood. We began by taking the practice as the unit of analysis and imputing average patient characteristics (such as average age or proportion of male patients) using 2006 census data from the surrounding community, defined at either the dissemination area (DA) or census tract (CT) level. We then compared these imputed averages to the data collected by the Comparison of Models of Primary Care (COMP-PC) project. This allowed us to test the accuracy and reliability of practice-centred census imputation against traditional data collection methods and ask whether the former is an adequate substitute for the latter.

METHODS

Design

The COMP-PC study was a cross-sectional study with a concurrent nested qualitative component in which several performance parameters were evaluated. Full details on the methodology of the project can be found in a separate publication.[10] The study was approved by the Ottawa Hospital Research Ethics Board. Data collection took place between October 2005 and June 2006.

Geocoding

Our census data were compiled at 2 levels of aggregation: DAs (400 to 800 individuals) and CTs (1000 to 3000 individuals). Postal codes for each practice were recorded and converted into DAs and CTs using the Statistics Canada 2006 Postal Code Conversion File and the "best single link" function.[10,11] Census data were subsequently collected from the Computing in the Humanities and Social Sciences Canadian Census Analyser.[12] Data on income were not available for many of the DAs owing to Statistics Canada's reporting policies for small sample sizes; thus, we did not analyze this variable at the DA level.

Sample

The COMP-PC project aimed to recruit 35 practices from each of the 4 most common primary care models in Ontario: community health centres (CHCs), fee for service (FFS) practices (including the newly formed family health groups), family health networks (FHNs), and health service organizations (HSOs). Participating practices were recruited from a sampling frame that included all known FHNs (N = 94), CHCs (N = 51), and HSOs (N = 65) in the province. The FFS sampling frame of 155 practices represented a random sample extracted from a province-wide list of 1884 practices. We excluded practices that did not offer comprehensive primary care services for adults or that had belonged to their respective models for less than 1 year, as well as practices in which less than 50% of the providers consented to participate. For this substudy we excluded rural practices, as rural areas share a single CT code and are thus impossible to differentiate at that level of aggregation.

Patient sampling within practices aimed to collect between 30 and 50 patient surveys per site. Patients were eligible if they were 18 years of age or older, were not acutely ill or cognitively impaired, and could communicate in English or French or through a translator. Practice-level averages were calculated from patient responses to these waiting-room surveys.

Outcome measures

Three descriptive statistics were calculated to summarize the relationship between census and practice data: mean absolute relative error (ARE), median ARE, and Spearman rank correlations. Statistical analyses were carried out using SPSS, version 17.0. We did not calculate *P* values for differences between practice and census data because our focus was on clinically important, rather than statistically significant, differences.

Mean ARE is the average unsigned difference between the estimated and true values divided by the true value. This measures the size of the errors as proportions of the variables being estimated (eg, a mean ARE of 0.5 indicates that the average error size is 50% of the true value). Formally, the mean ARE is calculated as follows:

$$\frac{1}{n}\sum_{i=1}^{n}\frac{|\text{ true value}_i - \text{estimated value}_i\,|}{\text{true value}_i}$$

Mean ARE is superior to the mean error as a measure of accuracy owing to its ability to distinguish offsetting positive and negative errors. Mean ARE also outperforms the mean squared error in its handling of outliers.[13] For calculation purposes, we took patient survey results as the "true" value, and census data as the estimator.

Median ARE was calculated to provide a summary of errors that are insensitive to outliers and to assess skewness through mean-median comparison.

Our final statistic was the Spearman rank correlation, computed between practice and census averages. The Spearman rank correlation measures the level of ordinal agreement between 2 variables, and approaches +1 or -1 as the variables become either monotonically increasing (+1) or decreasing (-1) functions of each other. Thus, it provides a different perspective on census-practice similarities. In particular, it is sensitive to relationships that cannot be detected using absolute relative errors. For example, suppose that large households were overrepresented in our waiting-room sample relative to census values (a very real possibility because, all other factors being equal, larger households have more members who might be sick at any given time). This would result in a high mean ARE. However, if the level of overrepresentation were systematic and stable (suppose an average of 0.5 household members), the relationship would still yield a high value for the Spearman rank correlation. This would then indicate that the 2 variables were systematically related, possibly allowing researchers to adjust the census value to correct for systematic errors and accurately estimate practice averages.

RESULTS

The final sample consisted of 28 CHC, 31 FFS, 26 FHN, and 31 HSO practices; 4413 surveys were collected from patients receiving care at these sites. Response rates to the patient survey ranged from 94% in FHNs to 77% in CHCs, with an overall response rate of 82%.

We compared the sociodemographic profiles of CHC patients surveyed with those of all CHC patients contained in the CHC practice electronic patient registry. Other practice models lacked a centralized patient registry that would have allowed us to make similar comparisons. Community health centres are the model for which our surveys are least likely to be representative, as CHCs serve marginalized populations with access barriers.[10] Thus, the representativeness of our CHC sample is a worst-case measure of overall survey representativeness. As anticipated, the waiting-room sample was older, poorer, and more likely to be female than the overall CHC population, reflecting the profile of those who make more use of primary care services (data not shown).[10]

As presented in **Tables 1** and **2**, we observed large and highly variable differences between census and practice means. In almost all cases, the median ARE was lower than the mean ARE. The overall mean ARE was high for almost all variables, with the exception of proportion of English speakers. Mean and median ARE were slightly larger for CHC practices than for non-CHC practices (data not shown). Although universally positive, Spearman rank correlations were generally low.

Table 1. Accuracy of dissemination area estimation of sociodemographic characteristics: *N* = *116 practices.*

VARIABLE	MEAN ARE	MEDIAN ARE	SPEARMAN RANK CORRELATION
Male sex	0.82	0.47	0.02
Aged 65 years and older	0.78	0.54	0.30
Unemployed	0.78	0.67	0.11
Average household size	0.27	0.26	0.36
English speakers	0.12	0.05	0.35
White	0.38	0.11	0.47
Without high school education	0.81	0.46	0.39
Not born in Canada	0.83	0.43	0.48

ARE—absolute relative error.

Table 2. Accuracy of census tract estimation of sociodemographic characteristics: *N = 116 practices.*

VARIABLE	MEAN ARE	MEDIAN ARE	SPEARMAN RANK CORRELATION
Male sex	0.82	0.49	0.04
Aged 65 years and older	0.64	0.41	0.17
Unemployed	0.79	0.61	0.15
Average household size	0.25	0.24	0.30
English speakers	0.12	0.49	0.32
White	0.39	0.10	0.37
Without high school education	0.90	0.42	0.22
Not born in Canada	1.36	0.55	0.32
Household income < $10 000	0.60	0.56	0.13
Household income $10 000-$49 999	0.59	0.36	0.12
Household income $50 000-$79 999	0.79	0.28	0.18
Household income ≥ $80 000	1.67	0.81	0.15

ARE—absolute relative error.

DISCUSSION

Researchers should not estimate the demographic profile of a practice using census data centred on the location of that practice, as the differences we observed were large enough to render census-based estimates invalid. The reasons for these differences could be methodological or substantive. Our findings support the view that the observed differences were substantive and systematic, rather than of methodological origin.

There are several potential methodological explanations for the differences we observed between census and practice averages. One is the difference in age cutoffs used by our data sources. Statistics Canada calculated unemployment data for individuals 15 years of age and older, whereas our survey samples were restricted to individuals 18 years of age and older. It is also possible that overrepresentation of frequent health care users (such as the elderly) biased our patient sample, although comparisons between patient survey data and chart audit data collected from the same practices found similar mean ARE and median ARE scores (data not shown). Because chart audits reflect the overall practice population, they are insensitive to more frequent use by certain individuals or groups; the fact that our results for chart audits and patient surveys align suggests that the patient surveys were not affected by overrepresentation problems. Our geocoding could also be at fault, given that we restricted census data to a single DA or CT, rather than averaging several areas surrounding the practice. Yet given the large size of the mean ARE for almost all variables, such explanations are unlikely to account for the differences observed. It seems more likely that we observed systematic differences because patient populations differ systematically from the practices' local communities.

There are a number of mechanisms that might give rise to such systematic differences, all of which imply that either practices draw many patients from outside the local area or that they draw their patients from the local population in an unrepresentative fashion. One explanation stems from the shortage of family physicians in Canada. Individuals often search areas far from their homes, and continue to visit the same family doctor even if they relocate. Alternatively, modern commuter lifestyles might also undermine attachment to residential neighbourhoods for access to services, leading individuals to choose practices closer to their workplaces or their children's schools. A further complicating factor is the existence of practices (particularly CHCs) that specialize in caring for populations defined by social, ethnolinguistic, or disease morbidity (eg, AIDS clinics) criteria, as these populations might not be concentrated geographically. Our patient survey captured only the first 3 digits of the patient's postal code, known as the forward sortation area (FSA). Forward sortation areas are used by Canada Post to deliver mail and do not map onto census divisions like CTs and DAs because a medium-sized city might be covered by only 1 FSA, yet have dozens of CTs and DAs. Thus, we cannot directly measure the proportion of patients who live outside the practices' CTs or DAs. However, we examined whether patients and clinics shared the same FSAs, and found that only 32% of patients lived in the same FSAs as their clinics. This implies that most patients live outside the geographic neighbourhoods of their primary care practices.

Regardless of what explains these differences, our results have clear policy implications for practice outreach programs and community-based approaches to primary care. According to our results, a practice's community, if it can be defined at all, is unlikely to be a geographically bounded entity. Furthermore, data collection about this community (such as health needs assessments) must be undertaken at the level of the practice itself, as an obvious source of outside data (the census) performed poorly as a substitute. Researchers and policy makers should take these facts into account when investigating issues relating to practice-community and practice-patient links.

Conclusion

This study investigated the accuracy of estimating patient characteristics using census data from Ontario. Our sample consisted of 116 practices located in non-rural areas and drew data from the 2006 census at 2 levels of aggregation: DAs and CTs. We found important differences between census data centred on the location of the practice and data generated by practice-level surveys. These differences were large enough to render practice-centred imputation invalid.

Mr Shortt is a law student at McGill University in Montreal, Que, and a student-at-law at Fasken Martineau Dumoulin. **Dr Hogg** is Director of Research and Professor in the Department of Family Medicine, Director of the C.T. Lamont Primary Health Care Research Centre at the Élisabeth Bruyère Research Institute, and Principal Scientist at the Institute of Population Health, all at the University of Ottawa in Ontario. **Dr Devlin** is Professor in the Department of Economics at the University of Ottawa. **Dr Russell** is Adjunct Professor at the C.T. Lamont Primary Health Care Research Centre. **Dr Muldoon** is a family physician at the Somerset West Community Health Centre in Ottawa.

Acknowledgment

We thank **Ralph A. Shaw** for his helpful input regarding census data extraction, and **Meltem Tuna** for her statistical advice. Funding for this research was provided by the Ontario Ministry of Health and Long-Term Care Primary Health Care Transition Fund. The views expressed in this report are the views of the authors and do not necessarily reflect those of the Ontario Ministry of Health and Long-Term Care.

Contributors

All authors contributed to the concept and design of the study; data gathering, analysis, and interpretation; and preparing the manuscript for submission.

Competing interests

None declared

Correspondence

Mr Michael Shortt, 23 King's Landing Private, Ottawa, ON K1S 5P8;
telephone 514 236-8831; e-mail **michael.shortt@mail.mcgill.ca**

References

1. Evans RG, Stoddart GL. Producing health, consuming health care. *Soc Sci Med* 1990;31(12):1347-63.
2. Connolly V, Unwin N, Sherriff P, Bilous R, Kelly W. Diabetes prevalence and socioeconomic status: a population based study showing increased prevalence of type 2 diabetes mellitus in deprived areas. *J Epidemiol Community Health* 1999;54(3):173-7.
3. Shaw JE, Chisholm DJ. 1: epidemiology and the prevention of type 2 diabetes and the metabolic syndrome. *Med J Aust* 2003;179(7):379-83. Erratum in: *Med J Aust* 2003;179(10):526.
4. Robbins JM, Vaccarino V, Zhang H, Kasl SV. Socioeconomic status and type 2 diabetes in African American and non-Hispanic white women and men: evidence from the Third National Health and Nutrition Examination Survey. *Am J Public Health* 2001;91(1):76-83.
5. Shortt SE, Shaw RA. Equity in Canadian health care: does socioeconomic status affect waiting times for elective surgery? *CMAJ* 2003;168(4):413-6.
6. Roos NP, Mustard CA. Variation in health and health care use by socioeconomic status in Winnipeg, Canada: does the system work well? Yes and no. *Milbank Q* 1997;75(1):89-111.
7. Alter DA, Naylor CD, Austin P, Tu JV. Effects of socioeconomic status on access to invasive cardiac procedures and on mortality after acute myocardial infarction. *N Engl J Med* 1999;341(18):1359-67.
8. Paszat LF, Mackillop WJ, Groome PA, Zhang-Salomons J, Schulze K, Holowaty E. Radiotherapy for breast cancer in Ontario: rate variation associated with region, age and income. *Clin Invest Med* 1998;21(3):125-34.
9. Scrivener G, Lloyd DC. Allocating census data to general practice populations: implications for study of prescribing variation at practice level. *BMJ* 1995;311(6998):163-5.
10. Dahrouge S, Hogg W, Russell G, Geneau R, Kristjansson E, Muldoon L, et al. The Comparison of Models of Primary Care in Ontario (COMP-PC) study: methodology of a multifaceted cross-sectional practice-based study. *Open Med* 2009;3(3):e149-64. Epub 2009 Sep 1.
11. *2001 Canadian census postal code conversion file.* Toronto, ON: University of Toronto CHASS Data Centre; 2007. Available from: **http://dc2.chass.utoronto.ca/census/2001_pccf_vjan07.html**. Accessed 2007 Aug 1.
12. *Canadian Census Analyser* [website]. Toronto, ON: University of Toronto CHASS Data Centre; 2007. Available from: **http://dc1.chass.utoronto.ca/census/index.html.** Accessed 2012 Feb 15.
13. Armstrong JS, Collopy F. Error measures for generalizing about forecasting methods: empirical comparisons. *Int J Forecast* 1992;8(1):69-80.

✶ ✶ ✶