International Journal of Health Geographics

BioMed Central

Methodology

Open Access

Accuracy of city postal code coordinates as a proxy for location of residence

C Jennifer D Bow[2,4], Nigel M Waters[2,4], Peter D Faris[4], Judy E Seidel[3,4], P Diane Galbraith[4], Merril L Knudtson[1,4], William A Ghali*[1,3,4] and the APPROACH Investigators

Address: [1]Department of Medicine, University of Calgary, 3330 Hospital Drive NW, Calgary, AB, T2N 4N1, Canada, [2]Department of Geography, University of Calgary, 2500 University Drive NW, Calgary, AB, T2N 1N4, Canada, [3]Department of Community Health Sciences, University of Calgary, 3330 Hospital Drive NW, Calgary, AB, T2N 4N1, Canada and [4]Centre for Health and Policy Studies, Faculty of Medicine, Department of Community Health Sciences, University of Calgary, 3330 Hospital Drive NW, Calgary, AB, T2N 4N1, Canada

Email: C Jennifer D Bow - jbow@chspr.ubc.ca; Nigel M Waters - nwaters@ucalgary.ca; Peter D Faris - faris@ucalgary.ca; Judy E Seidel - jseidel@ucalgary.ca; P Diane Galbraith - dgalbrai@ucalgary.ca; Merril L Knudtson - knudtson@shaw.ca; William A Ghali* - wghali@ucalgary.ca; the APPROACH Investigators -

* Corresponding author

Published: 18 March 2004

International Journal of Health Geographics 2004, **3**:5

Received: 01 December 2003
Accepted: 18 March 2004

This article is available from: http://www.ij-healthgeographics.com/content/3/1/5

Abstract

Background: Health studies sometimes rely on postal code location as a proxy for the location of residence. This study compares the postal code location to that of the street address using a database from the Alberta Provincial Project for Outcome Assessment in Coronary Heart Disease (APPROACH©). Cardiac catheterization cases in an urban Canadian City were used for calendar year 1999. We determined location in meters for both the address (using the City of Calgary Street Network File in ArcView 3.2) and postal code location (using Statistic Canada's Postal Code Conversion File).

Results: The distance between the two estimates of location for each case were measured and it was found that 87.9% of the postal code locations were within 200 meters of the true address location (straight line distances) and 96.5% were within 500 meters of the address location (straight line distances).

Conclusions: We conclude that postal code locations are a reasonably accurate proxy for address location. However, there may be research questions for which a more accurate description of location is required.

Background

Postal codes are often used in health research to define geographical location of residence either directly or via linkages to census geographical units [1-11]. The fields of epidemiology and medical geography both examine research questions where analysis of location of residence is either desirable or required. Examples of types of spatial analysis requiring exact location of residence include location allocation, cluster analysis, and point-pattern analysis. These analysis methods are often applied when

searching for geographical clusters of disease or geographical patterns of health service utilization.

Despite its widespread use in health research, the validity of the postal code for location of residence is not known. Two previous studies have addressed methodological issues regarding the use of postal code to define location in medical studies. Burra, Jerrett, et al. [12] identified error associated with postal code location and attributed some of this error to inaccuracies in the Statistics Canada postal code conversion file [13]. Glass, Gray, et al [14] assessed the validity of cancer registry data in Scotland noting a 44 percent error rate in the database including errors relating to the input of the postal code, which led to an error in their cluster analysis. While useful, the above two studies alone do not provide a clear indication of how valid location derived from postal codes is relative to location derived from street addresses.

The Alberta Provincial Project for Outcome Assessment in Coronary Heart Disease (APPROACH©) is a large, population-based, clinical registry that captures all patients undergoing catheterization and subsequent revascularization in Alberta, Canada [15]. The registry has been used to study outcomes and processes of care relating to cardiovascular disease and has recently been applied to an analysis of geographical location of residence as a predictor of cardiac care [16]. The APPROACH clinical registry contains both street address and postal code for each patient and provides an opportunity to assess the accuracy of postal code location as a proxy for residence location.

Accordingly, the objective of this methodological study was to determine the accuracy of postal code location compared to address location in the urban setting of a large Canadian City. We confined the analysis to urban locations because it is in urban locations that the postal code is likely to be most accurate and problems exist with postal code misclassification in rural areas. Precise geographical coordinates for street address locations were derived using a detailed street network file and were compared to the geographic coordinates derived for the postal code from the Statistics Canada Postal Code Conversion File.

Methods

Data sources

The APPROACH database is a clinical data collection initiative that began in 1995 that includes all patients undergoing cardiac catheterization in the province of Alberta, Canada. For this study the most important characteristics of the database were the patient's unique identifier, the address, and the postal code. The data for this study were confined to residents of the City of Calgary undergoing cardiac catheterization in the calendar year 1999.

The Canadian Postal Code is 6 characters long with alternating alphabetic and numeric characters of 'ANA NAN'. The first three characters identify a major urban or rural area known as the 'forward sortation area' and the last three characters identify the smallest delivery unit – the 'local delivery unit' [17]. The 'local delivery unit' may indicate a specific city block, a single building, or a large volume mail receiver [17].

The Postal Code Conversion File, created by Statistics Canada and Canada Post, contains information on each postal code in Canada. The Postal Code Conversion File provided the geographic coordinates of each postal code in latitude and longitude. Within major urban areas, postal code address ranges are linked to the National Geographic Base, and where possible, a block-face link and its coordinates are identified. Municipality maps and direct contact with local authorities are then used to derive precise locations for postal codes relative to these block face links [13].

The City of Calgary Street Network File, created by the City of Calgary, contains street information for the city, as recorded in 2001. The Street Network File permitted us to determine the geographic coordinates of each address recorded in the APPROACH database through interpolation of individual home addresses on a range of addresses represented by a street (see section on Determining Location below).

Study sample

We used APPROACH data from calendar year 1999 for this study, and had access to detailed street network data for the City of Calgary. Records of patients from rural areas, Rural Routes, rural sites, and cities other than Calgary were removed from the APPROACH dataset. We also excluded patients who had post office box addresses, because such addresses can not be geocoded, unless ancillary sources of data – not readily available to us – are used (e.g., post office box rental records, department of motor vehicle records, 911 emergency service records) [18]. A total of 3180 APPROACH cardiac catheterization cases were thus screened, and among these 177 were excluded on the basis of having non-Calgary addresses recorded. Another 56 cases were excluded because their address was recorded as a Rural Route, site, or as a post office box number. This left 2947 cases for subsequent evaluation. Input errors in the dataset were identified among these 2947 cases by screening for mismatches between address and postal code. These errors were corrected, where possible, by comparing the address and postal code fields recorded in the APPROACH registry to the corresponding address and postal code fields recorded in the City of Calgary 2001 Phone Book. Among the 2947 APPROACH registry records studied, 481 (16.3%) contained such

mismatch errors. The major reasons for these discrepancies were: postal code and address did not match, missing or incomplete address, address did not exist, wrong street name or number, consequential spelling error, or the postal code was recorded with one or more errors in its letters or digits (errors that presumably arise from the complexity of manually keying in postal codes and street addresses). Of the 481 mismatch errors detected in the database, 268 records were corrected and the remaining 213 uncorrected cases had to be excluded from further analysis. An additional 7 cases were removed during postal code conversion. Subsequent analysis focused on the remaining 2727 records with complete and accurate information on both the address and postal code.

Determining location

The address location was determined spatially by a process in a Geographical Information System (GIS) known as geocoding. Because each Street Network File contains information for each arc or street, including address ranges, it is possible to compute where specific addresses would be hypothetically located on that arc or street. Arcs are divided into equal segments that permit individual address points to be estimated at an appropriate place on the street network. Geocoding was used in ArcView 3.2 GIS Software http://www.esri.com to determine the spatial location of each address in the APPROACH database. Of the 2727 addresses available, 2687 (98.5 percent) of the address locations were geocoded by ArcView. Addresses that did not geocode were likely due to missing address information in the City of Calgary street network file database or, less likely, may have represented a missed error in the APPROACH database. The address locations were then mapped and the location of the x and y coordinates were found to the nearest meter that correspond to the latitude and longitude of the North American Datum of 1927, Zone 11.

The postal code spatial location was determined by cross-referencing the postal code in the APPROACH database to the geographic coordinates found in Statistic Canada's Postal Code Conversion File. This cross-referencing was performed using Microsoft Access's http://www.microsoft.com query function. Once the postal code in the APPROACH database was matched to a geographic coordinate it was possible to map the location of each of the postal codes in ArcView 3.2. Only the 2687 matching pairs of the 2727 address points were used. These matched pairs were those for which a precise location could be determined based on street address, as described in the preceding paragraph. The postal code locations were found to the nearest meter that correspond to the latitude and longitude of the North American Datum of 1927, Zone 11.

Data analysis

The address and postal code spatial locations were compared on a map to determine visually how closely the postal code represented the address. The distance between each of the address and postal code location pairs was then determined. The x (east-west) and y (north-south) dimensions of the differences between paired points was computed for each pair of points and a scattergram of these differences was created in the Statistical Package for the Social Sciences (SPSS, http://www.spss.com). Straight-line (Euclidean) distances were calculated using the formula $z = \sqrt{((x_i - x_j)^2 + (y_i - y_j)^2)}$, and city-block (Manhattan) distances were calculated using the formula $z = (x_i - x_j) + (y_i - y_j)$. The latter were studied because they approximate the true travel distance between any two given points in typical urban grids where buildings often prevent straight line travel. Descriptive statistics were computed for the Euclidean and Manhattan distances using SPSS.

Results

The postal code locations for the City of Calgary are displayed spatially in Figure 1. Mapping the postal code location is a visual tool that can help in interpretation of geographical relationships in health data, provided, of course, that the postal code location presented in Figure 1 is reasonably accurate. In Figure 1, the size of the marker representing each postal code is varied to indicate the number of study subjects (1 to 10) drawn from that postal code.

The address and the postal code locations for a representative area of the City of Calgary is shown in Figure 2. In this figure, one can see how the distances between the address and postal code locations can vary across pairs. The range of location differences in this example of a small downtown area is between 15.8 meters and 85.5 meters. The range may be larger or smaller in other areas of the City depending on the road type and length, and on the size and shape of the postal code area in question.

The x and the y-axis of the address and postal code location differences are plotted in Figure 3. Scattergrams are displayed in this figure at various scales. The full-extent plot displays a number of outliers where the spatial distance between the address and the postal code can be more than 4000 meters. For the majority of data pairs however, the distance between the address and the postal code locations is less than 50 meters.

The address and postal code locations compare the x and y-axis differences, Euclidean distance, and Manhattan Distance in Table 1. The means of the absolute x and y differences were 95.7 and 89.9 with standard deviations of 271.1 and 259.0 respectively. The Euclidean (straight-

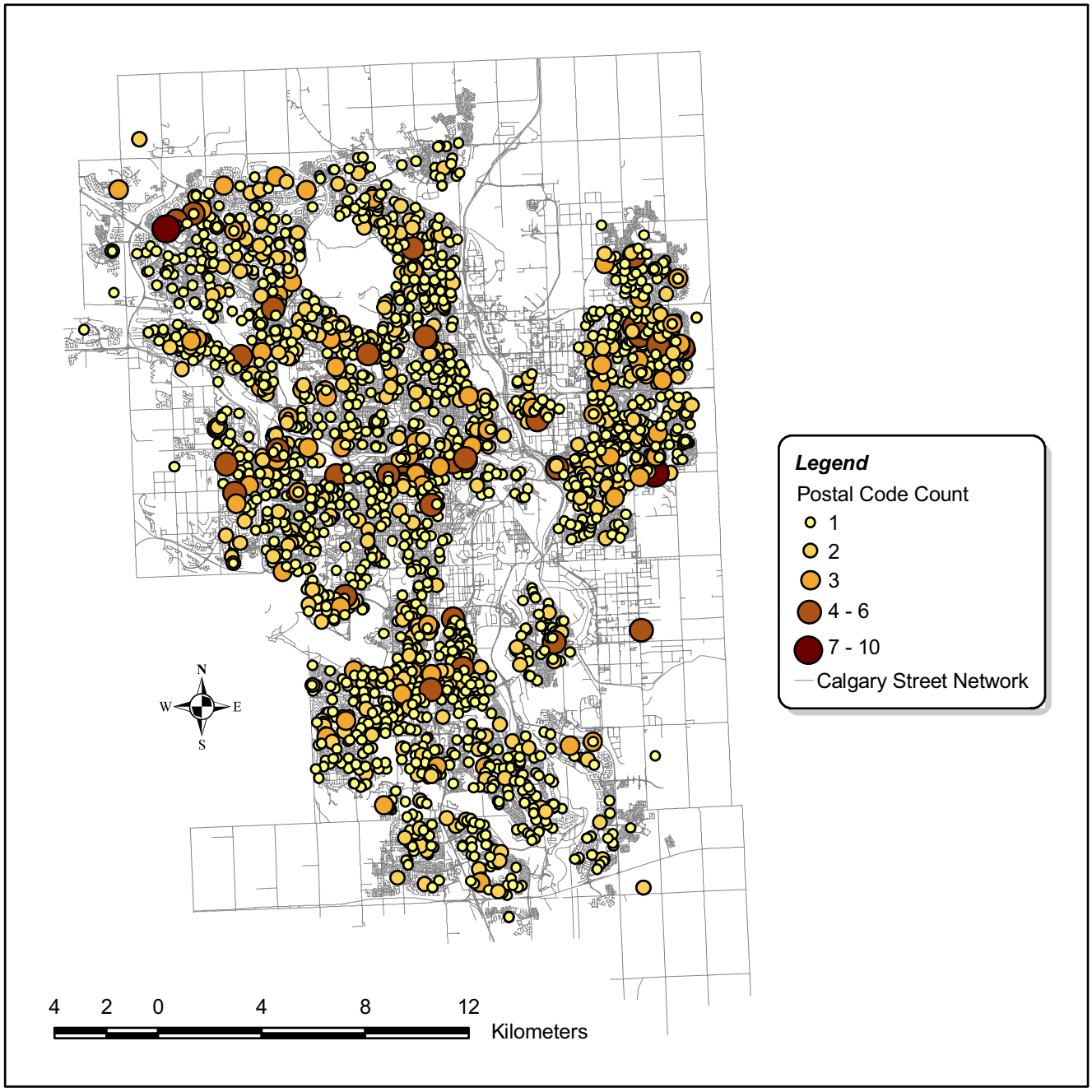

Figure 1
City of Calgary postal code locations

line) distance has a higher mean of 146.2, a greater range of values, and a higher standard deviation of 369.3. As expected, Manhattan distances were larger than the absolute x and y values and the Euclidean distance because Manhattan distance attempts to describe the restrictive movement of travel in typical urban centers, which use a rectangular grid pattern of streets. By assessing the 25th percentile, the median, and the 75th percentile of the x distance, y distance, the Euclidean distance, and the Manhattan Distance, and by comparing these values to the corresponding means and standard deviations shown in Table 1, it becomes apparent that the distributions of distances is skewed, with a few observations demonstrating large distance differences.

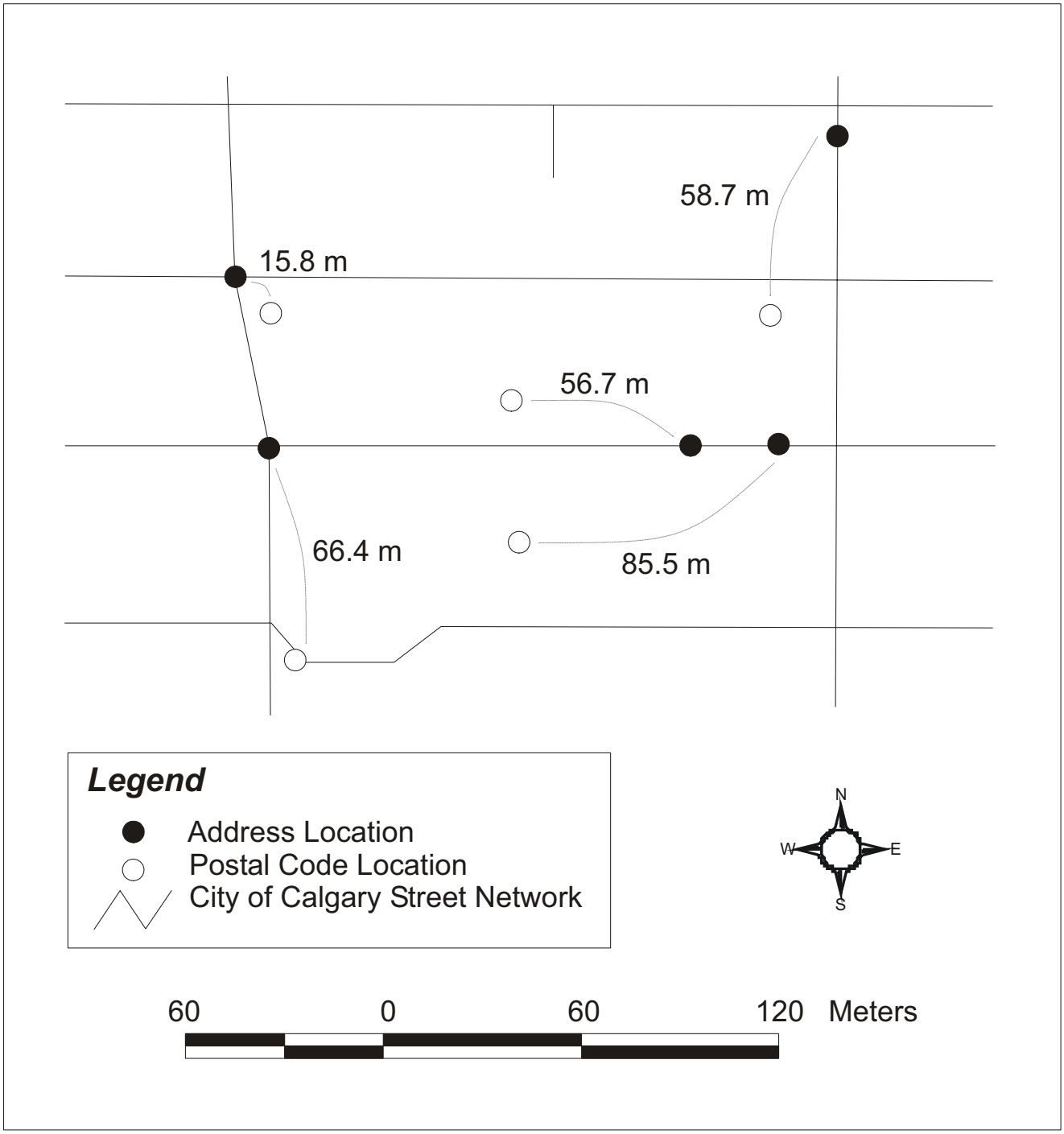

Figure 2
Example of a small area in the City of Calgary indicating address and postal code locations

The proportion of patients falling into various distance difference categories is shown in Table 2. For Euclidean distance, 34.4 percent of the postal code locations are within 50 meters of the address location, 65.8 percent are within 100 meters, and 96.5 percent are within 500 meters. To place these values in perspective, if one city

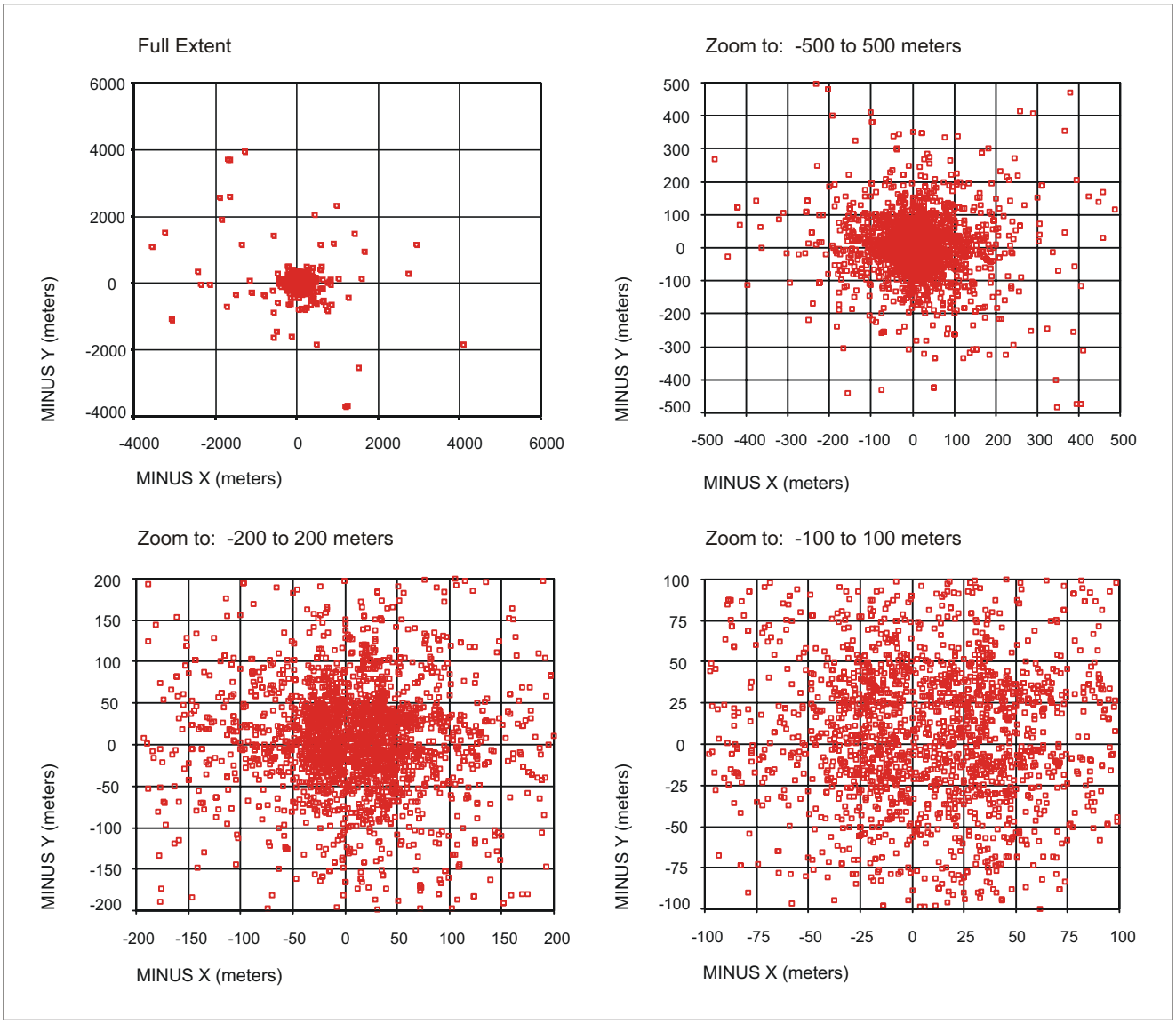

Figure 3
Scattergrams of the City of Calgary x and y distances from the full extent to zoomed in versions at 500, 200, and 100 meters

Table 1: Descriptive statistics of distance difference between pairs of addresses and postal codes (meters)

Statistic	X distance	Y distance	Euclidean distance	Manhattan distance
Range	4092.1	3928.7	4480.9	5919.4
Minimum	.02	.01	1.51	2.05
Maximum	4092.09	3928.66	4482.38	5921.45
25th Percentile	20.1	18.6	41.5	52.7
75th Percentile	85.6	81.6	128.2	158.3
Median	38.4	38.3	69.1	87.2
Mean	95.7	89.9	146.2	185.6
Standard Deviation	271.1	259.0	369.3	474.6

Table 2: Distance differences between pairs of addresses and postal codes in specific ranges

Distance	Euclidean distance			Manhattan distance		
	N	% in Category	Cumulative %	N	% in Category	Cumulative %
0 to 50 m	923	34.4	34.4	590	22.0	22.0
51 to 100 m	846	31.5	65.9	916	34.1	56.1
101 to 200 m	592	22.0	87.9	710	26.4	82.5
201 to 500 m	232	8.6	96.5	355	13.2	95.7
Over 501 m	94	3.5	100.0	116	4.3	100.0

block is approximately 200 meters, then 87.9 percent of the postal code locations are within one typical city block of the address location for straight-line distances. For Manhattan distance, 22.0 percent of the postal code locations are within 50 meters of the address location, 56.1 percent are within 100 meters, and 95.7 percent are within 500 meters. Accordingly, 82.5 percent of the postal code locations are within one typical city block of the address location for Manhattan distances.

Conclusions

Using the postal code and address information from the APPROACH database, we were able to determine the spatial location of the street address and postal code and the difference in distance between those paired points. The major finding of the present study is that the postal code location closely approximates the location of residence for a large majority of patients with over 80 percent of the postal codes being within 200 meters of the address, and approximately 95 percent of the postal codes being with 500 meters of the street address.

How close is close enough for studies in epidemiology and medical geography? The answer to this question depends on the amount of accuracy needed to answer a specific research question, which in turn depends on the nature of the research question itself. For example, investigation of a regional outbreak of E. coli gastrointestinal disease would probably not be hampered by a random misclassification of locations of 200 meters or less for the majority of cases. In contrast, the planning of fire station or emergency medical service (EMS) networks within city limits is an example of a scenario where the accuracy of postal code locations demonstrated in our study may not suffice. Indeed, dealing with only approximately 80 percent accuracy rates for placing residences within a city block might lead to problems if one were to rely on postal codes only in the planning of locations for EMS and fire stations. The consequence of using postal codes would potentially include misclassification of events and eventually sub-optimal response times and access to service. Accordingly, researchers and policy makers need to consider carefully the geographical issue that they are studying and should use postal codes only if an approximate localization of individuals will suffice. For our research on localization of patients undergoing cardiac catheterization to assess overall geographic equity and access within the City of Calgary, we conclude that the spatial relationships derived from postal codes (and presented in Figure 1) represent a reasonable proxy for the true location of the residence. We are certain, however, that there are geographical scenarios for which more accurate determinations of location are required.

While a large majority of postal code locations were within 500 meters of true location of residence, it is notable that some postal code locations were large distances from the actual street address location (up to 4,482 straight-line distance meters away). The reasons for these highly discrepant cases are unclear, but could include errors in the postal code conversion file geographical coordinates (something that has been anecdotally reported by researchers who work with this file [1,2], and [12]), errors in the geocoding process in ArcView 3.2 for the street address, or a missed error in the cleaning process of the APPROACH database. Of course, it is possible that some of these larger differences represent true differences between street address and postal code locations.

The major limitation of the present study is that it focuses only on a single large urban Canadian city. Accordingly, the results do not necessarily apply to other Canadian cities or to cities in other countries. However, there are many reasons to suspect that the findings for the single city studied apply to other Canadian cities, and perhaps also to urban settings in other countries. In contrast, however, the results are unlikely to apply to rural postal codes where single postal codes cover large geographic areas. It is thus clear that additional research is needed to address the locational accuracy of postal codes in other environments. A second caveat to our findings is that we consider address locations geocoded from street network files as the 'refer-

ence standard' against which postal code location is compared. Such address locations are, in fact, themselves merely an estimate of location that may or may not be entirely accurate. Of relevance in this regard, Bonner and colleagues [19] recently compared geocoded address locations with true 'gold standard' location measures derived from Global Positioning System satellite receivers, and demonstrated slight discrepancies between the two. Another important caveat to our findings is that we made corrections to the discrepant address and postal code pairings whenever possible. The result of these corrections is likely to be an overall improvement in accuracy of location that would not otherwise occur if real addresses had not been available. This means that our findings from APPROACH may slightly overestimate the accuracy of postal code locations in 'typical' health databases. A final caveat (that may bias findings in the opposite direction) is that we relied on the Statistics Canada Postal Code Conversion File to determine the latitude and longitude of postal code locations. This is a widely used approach. However, anecdotal reports suggest that the latitude and longitude values in Statistics Canada data may be somewhat erroneous due to a geographical projection problem in the Statistics Canada street network file. These projection problems may in turn adversely affect the accuracy of postal code locations.

Despite these limitations and caveats, our study is informative. Our findings indicate that, although postal code location is not a perfect representation of street address location, the estimate is very close for a majority of cases. Researchers and policymakers interested in conducting and interpreting results of epidemiological or geographical studies need to consider carefully, on a case-by-case basis, whether a misplacement of 200 to 300 meters (or more) in spatial location is problematic to the objectives of their analysis. If the misplacement is not a major concern and given the random nature of that misplacement, the postal code can be used for analysis. In other instances, more precise information on address location should be obtained and used in analysis.

Authors' contributions

CJDB compiled data, analyzed data, and wrote paper; NMW provided oversight of geographic analysis and provided comments on drafts of paper; PDF assisted with data compilation and provided comments on drafts of paper; JES provided input into interpretation and provided comments on drafts of papers; PDG assisted with data compilation and provided comments on drafts of paper; MLK principal investigator of APPROACH who oversees entire initiative and approved paper for submission; WAG supervised all aspects of this project and contributed to manuscript writing and editing.

Acknowledgements

The APPROACH initiative was initially funded in 1995 by a grant from W. Garfield Weston Foundation. The ongoing operation of the project has been made possible by contributions from the Province-Wide Services Committee of Alberta Health and Wellness, Merck Frosst Canada Inc., Monsanto Canada Inc. – Searle, Eli Lilly Canada Inc., Guidant Corporation, Boston Scientific Ltd, Hoffmann-La Roche Ltd, and Johnson & Johnson Inc – Cordis. We appreciate the assistance of the Calgary Health Region and the Capital Health Authority in supporting on-line data entry by cardiac catheterization laboratory personnel.

We appreciate the assistance of Dr. Michelle Graham and Dr. Brent Mitchell for comments and revision suggestions.

Dr. Ghali is supported by a Health Scholar Award from the Alberta Heritage Foundation for Medical Research, Edmonton, Alberta, and by a Government of Canada Research Chair in Health Services Research.

References

1. Ng E, Wilkins R, Perras A: **How Far Is It to the Nearest Hospital? Calculating Distances Using the Statistics Canada Postal Code Conversion File.** *Health Rep* 1993, **5:**179-183.
2. Mackillop WJ, Zhang-Salomons J, Groome PA, Pazat L, Holowaty E: **Socioeconomic Status and Cancer Survival in Ontario.** *J Clin Oncol* 1997, **15:**1680-1689.
3. Spasoff RA, Gilkes DT: **Up-to-date denominators: evaluation of taxation family for public health planning.** *Can J Public Health* 1994, **85:**413-417.
4. Demissie K, Hanley JA, Menzies D, Joseph L, Ernst P: **Agreement in measuring socio-economic status: area-based versus individual measures.** *Chronic Dis Can* 2000, **21:**1-7.
5. Prince MI, Chetwynd A, Diggle P, Jarner M, Metcalf JV, James OFW: **The geographical distribution of primary biliary cirrhosis in a well-defined cohort.** *Hepatology* 2001, **34:**1083-1088.
6. Cousens SN, Linsell L, Smith PG, Chandrakumar J, Wilesmith JW, Knight RSG, Zeidler M, Stewart G, Will RG: **Geographical distribution of variant CJD in the UK (excluding Northern Ireland).** *Lancet* 1999, **353:**18-21.
7. O'Neill TW, Cooper C, Finn JD, Lunt M, Purdie D, Reid DM, Rowe R, Woolf AD, Wallace WA: **Incidence of distal forearm fracture in British men and women.** *Osteoporos Int* 2001, **12:**555-558.
8. Guernsey JR, Dewar R, Weerasinghe S, Kirkland S, Veugelers PJ: **Incidence of cancer in Sydney and Cape Breton County, Nova Scotia 1979–1997.** *Can J Public Health* 2000, **91:**285-292.
9. Mitchell JD, Gibson HN, Gatrell A: **Amyotrophic lateral sclerosis in Lancashire and South Cumbria, England, 1976–1986, A geographical study.** *Arch Neurol* 1990, **47:**875-880.
10. Smart RG, Adlaf EM, Walsh GW: **Neighbourhood socio-economic factors in relation to student drug use and programs.** *J Child Adolesc Subst Abuse* 1994, **31:**37-46.
11. Harris JM, Cullinan P, McDonald JC: **Occupational Distribution and Geographic Clustering of Deaths Certified to be Cryptogenic Fibrosing Alveolitis in England and Wales.** *Chest* 2001, **119:**428-433.
12. Burra T, Jerrett M, Burnett RT, Anderson M: **Conceptual and practical issues in the detection of local disease clusters: a study of morality in Hamilton, Ontario.** *Can Geogr* 2002, **46:**160-171.
13. Statistics Canada: *Statistics Canada Postal Code Conversion File September 2002 Postal Codes Reference Guide Catalogue 92F0153GIE: Ottawa, ON: Ministry of Industry*; 2002.
14. Glass S, Gray M, Eden OB, Hann I: **Scottish validation study of Cancer Registration data childhood leukemia 1968–1981.** *Leuk Res* 1987, **11:**881-885.
15. Ghali WA, Knudtson ML: **Overview of the Alberta Provincial Project for Outcome Assessment in Coronary Heart Disease. On behalf of the APPROACH investigators.** *Can J Cardiol* 2000, **16:**1225-1230.
16. Seidel JE, Ghali WA, Faris PD, Bow CJD, Waters NM, Graham MM *et al.*: **Distance and access to cardiac revascularization.** *Can J Cardiol* 2002, **Suppl B:**184B.
17. **Canada Postal Guide and Reference Tools** [http://www.canadapost.ca/business/tools/pg/manual/b02-e.asp]

18. Hurley SE, Saunders TM, Nivas R, Hertz A, Reynolds P: **Post office box addresses: a challenge for geographic information system-based studies.** *Epidemiology* 2003, **14:**386-391.
19. Bonner MR, Han D, Nie J, Rogerson P, Vena JE, Freudenheim JL: **Positional accuracy of geocoded addresses in epidemiologic research.** *Epidemiology* 2003, **14:**408-412.