Published in final edited form as:

Med Care. 2012 January ; 50(1): 99–106. doi:10.1097/MLR.0b013e31822944d1.

Estimating Proximity to Care: Are straight line and zipcode centroid distances acceptable proxy measures?

Robin L Bliss, PhD¹, Jeffrey N Katz, MD, MSc^{1,2}, Elizabeth A Wright, PhD¹, and Elena Losina, PhD^{1,3}

¹Orthopedic and Arthritis Center for Outcomes Research, Department of Orthopedic Surgery, Brigham and Women's Hospital, Harvard Medical School, Boston, MA

²Division of Rheumatology, Immunology and Allergy, Brigham and Women's Hospital, Harvard Medical School, Boston, MA

³Department of Biostatistics, Boston University School of Public Health, Boston, MA

Abstract

Background—Spatial accessibility of health care may be measured by proximity of patient residence to health services, typically in driving distance or driving time. Precise driving distances and times are rarely available. While straight line distances between zipcode centroids and between precise address locations are used as proxy measures for distance to care, the accuracy of these measures has received little study.

Methods—Among a cohort of Medicare beneficiaries, actual driving distances and times between patient residence and clinic were obtained from commercial software (MapQuest). We used a split-sample design to build and validate linear regression models that predict *actual* driving distances and times from *estimated* distances between zipcode centroids and between precise residential and hospital locations, adjusting for urban/suburban/rural residential status.

Results—On average, predicted driving distances and times were larger than actual values. Zipcode centroid distances alone predicted longer driving distances than observed values: rural +19% (3.2miles), suburban +23% (3.7miles), and urban +27% (2.0miles). Predicted time was 36% (9.4minutes) longer in rural, 32% (6.8minutes) longer in suburban, and 38% (4.7minutes) longer in urban areas than observed values. Including urban/suburban/rural categorization of residence improved the accuracy of predicted driving distance and time for suburban and urban areas but diminished accuracy for rural areas. Similar trends were observed for distance estimates from precise locations.

Conclusions—Distances between zipcode centroids and precise residential/hospital locations provide reasonable estimates of driving distance and time for epidemiologic research. Estimates are improved for suburban and urban residences when data are augmented by urban categorization.

Keywords

Estimating distance; health care accessibility; driving distance; driving time

Corresponding Author: Robin L Bliss One Brigham Circle - 4th Floor, Suite 16 Brigham and Women's Hospital 75 Francis Street Boston, MA 02115 rbliss@partners.org Phone: 617-525-8532 Fax: 617-525-7900.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Introduction

Access to health services may be conceptualized in two stages: the potential for and the actualized delivery of health care (1). An integral part of the potential for care is the distance in space and time between patient residence and health care centers, with greater distance becoming a potential barrier to care (1). Investigators have defined spatial health care accessibility as the distance or travel time between the locations of the patient's residence and of health care service receipt (2–6). In some studies, investigators have used the straight line distance between precise locations as a proxy measure for driving distance and driving time (2, 7). In studies where exact addresses were unavailable, investigators have approximated distance to health care as the straight line distance between residential and hospital zipcode centroids (3–6).

Spatial data may be categorized as point or aggregate data. Point data include an exact measure of geographic location such as street address or longitude and latitude of the location of interest, providing precise spatial measures. Statistics using point data can provide tests with improved power, sensitivity, and specificity (8–10) and the availability of individual-level data can reduce spatial confounding (11). Aggregate data are expressed on a community, census tract, or zipcode level where information is summarized across an area of residence. Aggregate data can often be obtained from disease registries or from the U.S. Census Bureau (12). These data provide greater privacy protection for subjects (13) and are less costly to obtain than point data (12). For residential location, an individual's residence may be expressed with precision using an exact street address (point data) or may be estimated using the zipcode centroid (aggregate data). Using the aggregate approximation, all subjects living in the same zipcode share the centroid as their residential location.

The accuracy of straight line distances between precise locations and between zipcode centroids has received little study. In this study, we examine the accuracy of straight line distances between locations (point data) and between zipcode centroids (aggregate data) in predicting driving distance and driving time using point and aggregate data collected from a Medicare cohort of total knee replacement (TKR) recipients.

Methods

Sample Description

Data were obtained from a retrospective cohort of Medicare beneficiaries who underwent elective primary TKR in the year 2000. A stratified random sample of all Medicare beneficiaries who underwent TKR in 2000 and were residing in Illinois, North Carolina, Ohio, and Tennessee was drawn by first sampling hospitals, stratified by the number of TKR operations performed per year, with probabilities proportional to annual TKR volume. Subsequently, subjects were sampled from hospitals with numbers of subjects per hospital varying between TKR volume strata (14).

Measures of Distance

We defined criterion standards for driving distance and driving time measurements between residence and hospital to allow an evaluation of the accuracy of estimates based on the straight line distances between precise locations and zipcode centroids.

Criterion standards: Driving distance and driving time between street addresses were computed using an automated procedure supplied through the MapQuest Developer Network (15). In its procedure, MapQuest applies Dijkstra's shortest path algorithm to recent road maps to plot the shortest driving route between the specified locations. The distance

traveled on each road is summed to obtain the driving distance between points (16). Driving time is calculated by MapQuest as the quotient of the total miles by the average speed limit (miles per hour) for each road in the route (17).

In some instances, MapQuest software is unable to determine the exact location due to changes in roadways over time. This produces variability in the automated procedure results. For quality control, we performed a manual review of driving distance and time for a randomly selected 2% (n=30) of all observations. For this review, we manually entered subject residential and hospital addresses into MapQuest and recorded the driving distance and driving time. While 5 of the points (16.7%) had observed values differing by at least 25% from the manually verified values, only two of these differences had magnitudes greater than one mile. None of the values was corrected.

In a second, independent validation procedure, we identified all observations with driving distances at least one mile shorter than the corresponding straight line distances between points. These events are implausible, indicating error. Nine of the 1,135 observations (0.79%) had such erroneous distances and were corrected using driving distances and times observed from manual review.

Estimating distances: We considered two estimates of distance: straight line distance between patient residence and hospital locations and straight line distance between zipcode centroids. Both were calculated using the Great Circle Distance Formula:

$$D=3963.0 (\arccos [\sin (T_1) * \sin (T_2) + \cos (T_1) * \cos (T_2) * \cos (G_2G_1)]),$$

where T_i is the latitude and G_i is the longitude of locations 1 and 2 in radians.

Factors related to accuracy of estimation: We considered two factors possibly related to the accuracy of distance estimation: urban/suburban/rural categorization of residential neighborhood and state of residence. Urban, suburban, and rural categories were defined using data from the 2000 U.S. Census. Residences were defined as urban if 100% of the population in the respective census tract lived in urban areas. Suburban residences consisted of areas in which between 80 and 99% of the census tract population lived in urban areas, while rural residences had less than 80% of the population living in urban areas. Subjects lived in one of four states: Illinois, North Carolina, Ohio, and Tennessee.

Model Building and Validation

To facilitate the validation of prediction models for distances and time, we used a split-sample design by randomly separating the sample into four datasets: Training, Testing 1, Testing 2, and Testing 3 (Table 1). Using the Training data we produced descriptive statistics, computed correlations between measures of distance and time, and generated predictive models. The three Testing datasets were used to validate the estimates with independent data. Hypothesis tests were applied with a 0.05 significance level. All statistical analyses were performed using SAS Version 9.2 for Windows (18).

Building Prediction Models—We developed four sets of three models, two sets predicting MapQuest driving distance and two sets predicting MapQuest driving time between points. Predictors of interest included the following: a) the straight line distance between precise residence and hospital locations, b) straight line distance between residential and hospital zipcode centroids, c) urban, suburban, or rural categorization of residential census tract, and d) state of residence. (Models are described in Table 2.) We tested for two-way interactions between distance measures and urban categorization and

between distance measures and state of residence. Interactions were included in models if their p-values were statistically significant at the 0.05 level.

Predicting driving distance and time by straight line distance between residence and hospital locations: We produced three models to predict MapQuest driving distance and driving time between patient residence and hospital locations. The models had the following independent variables: Model I) straight line distance between points; Model II) distance between points and urban categorization of census tract; and Model III) distance between points, urban categorization, and state of residence. When statistically significant, twoway interaction terms were included in the models.

Predicting driving distance and time by straight line distance between zipcode centroids: Using the same model building approach as applied above, we predicted driving distance and driving time by: Model I) straight line zipcode centroid distance; Model II) zipcode centroid distance and urban categorization; Model III) zipcode centroid distance, urban categorization, and state of residence. When appropriate, interaction terms were included in the models.

Goodness-of-fit of predictive models: Adjusted model R^2 and Akaike Information Criterion (AIC) statistics were recorded. The Adjusted R^2 statistic is a version of the correlation coefficient (R^2), penalized for model complexity, as measured by increases in numbers of independent predictors. It ranges from 0 to 1 with larger values preferred. The AIC statistic is a goodness-of-fit value that balances model bias versus variability. It rewards models with small residuals (the difference between the observed and fitted values) but penalizes complex models. The AIC statistic can range between 0 and positive infinity with smaller values indicating better model fit (19). Within each model building scheme, R^2 and AIC statistics were compared to select models that best predicted MapQuest driving distance and driving time in the Training data.

Model Validation—Models attaining high correlations (high adjusted R^2 , low AIC statistics) between observed and estimated distances were selected for validation in the three Testing datasets. Predicted MapQuest driving distances and driving times were derived by fitting the regression equations produced from the Training data to the Testing datasets. Models were applied to evaluate the accuracy of predicted driving distance and time when compared to observed driving distance and time in the three Testing datasets. Models were also applied to stratified data to determine whether the prediction equations may be generalized across states and urban, suburban, and rural observations. Model fit was evaluated using the unadjusted model R^2 statistic and accuracy of the estimates was

evaluated by calculating the mean relative bias, $100 \frac{\text{observedvalue} - \text{estimatedvalue}}{\text{observedvalue}} \%$.

This research was approved by the Institutional Review Board at Brigham and Women's Hospital.

Results

Sample Characteristics

Addresses from 1,135 TKR subjects and the corresponding hospitals where they had TKR were geocoded to compute measures of distance between subject residence and hospital location. Thirty-two percent of subjects were from Illinois, 18% from North Carolina, 34% from Ohio, and 16% from Tennessee. Overall, subject residences were 28% urban, 21% suburban, and 51% rural. Illinois had the most subjects living in urban areas (50%), while

North Carolina had the most living in rural areas (78%). Each of the Training and Testing datasets were composed of between 273 and 292 randomly assigned subjects (Table 1).

In the Training data, the average straight line zipcode centroid distance (Mean (SD): 10.0 miles (14.7)) was slightly shorter than the distance between straight line residential and hospital locations (Mean (SD): 10.3 miles (14.4)). The average MapQuest driving distance was around 14 miles and the driving time was near 21 minutes, corresponding to an average driving speed of 40 miles per hour (Table 3).

Distance and Time Estimates

The Pearson correlation between straight line distance between residence and hospital and MapQuest driving distance was 0.986 ($p < 0.0001$). The correlation between residence/hospital distance and driving time was slightly lower ($r = 0.951$, $p < 0.0001$). Similar correlations were observed between distance between zipcode centroids and driving distance ($r = 0.981$, $p < 0.0001$) and driving time ($r = 0.945$, $p < 0.0001$; Table 3).

Prediction Models for Driving Distance

Predicting driving distance by straight line distance between residential and hospital locations: Comparing model fits, adjusted R^2 and AIC statistics were very similar across Models I, II, and III (adjusted R^2 range: 0.972–0.973, AIC range: 1421.7–1425.7; Table 4). We found no significant interaction between urban categorization and straight line distance ($p = 0.7058$), indicating that the association between straight line distance and driving distance did not differ across urban, suburban, and rural strata. We observed a statistically significant interaction in Model III between distance and state of residence ($p = 0.0016$). In North Carolina, one straight line mile corresponded to a longer driving distance (1.27 miles) than in Ohio (1.16 miles; Table 4).

The most parsimonious model (Model I), was selected for validation among the Testing data. When the model was applied to the Testing data, Model I unadjusted R^2 values ranged from 0.973 to 0.998 indicating high correlation between the observed and estimated driving distances. Among data stratified by urban categorization or by state, all R^2 values were at least 0.911 (Table 5).

Predicting driving distance by straight line distance between zipcode centroids: Predicting driving distance by distance between zipcode centroids, all adjusted R^2 values were greater than 0.960. The minimal AIC statistic was observed for Model III (AIC=1510.0) including all predictors, closely followed by that of Model II (AIC=1511.8), including only zipcode distances and urban categorization (Table 4).

In Model II there was no significant interaction between distance between zipcode centroids and urban categorization ($p = 0.5125$). In Model III, a significant interaction between state of residence and distance between zipcode centroids ($p = 0.0061$) was observed where one mile in zipcode centroid distance corresponded to greater driving distances in Illinois (1.25 miles) and North Carolina (1.21) than in Ohio (1.13 miles) or Tennessee (1.13 miles; Table 4).

Model II provided a simple extension to the zipcode only model and was selected for validation using the Testing data. Model II unadjusted R^2 values from models predicting observed from estimated distance measures were greater than 0.940. Similar results were observed when models were applied to Testing data stratified by state and by urban categorization (Table 5).

Prediction Models for Driving Time

Predicting driving time by straight line distance between residential and hospital

locations: Examining model fits, all adjusted R^2 values were at least 0.904, with the highest observed for Model III (adjusted $R^2=0.921$). AIC statistics indicated Model III to have the greatest model fit, with Model II following (Model III AIC=1866.5, Model II AIC=1887.0; Table 4).

Significant interactions were observed between residence/hospital distance and urban categorization ($p<0.0001$) and between residence/hospital distance and state of residence ($p<0.0001$) indicating that associations between distance and driving time differed by both urban categorization and by state. Both Models II and III included interaction terms.

Though Model III provided a slightly better fit to the data, the goodness-of-fit statistics were similar to one another and we applied Model II to Testing data because it was more parsimonious. All Model II unadjusted R^2 values were at least 0.915 when applied to Testing data and the minimal R^2 value observed for stratified data was 0.822 indicating strong model fit across urban and state strata (Table 5).

Predicting driving time by straight line distance between zip code centroids:

All adjusted R^2 values were at least 0.890 and AIC statistics ranged from 1899.3 for Model III to 1941.8 for Model I. Though the difference in AIC statistics was nearly 20 units (Model II AIC=1917.4, Model III AIC=1899.3), the difference from Model II to Model III adjusted R^2 values was only 0.007 (Model II $R^2=0.904$, Model III $R^2=0.911$) indicating similar fit between the models (Table 4).

As was observed for residence/hospital distances, when models were applied to predict driving time from distance between zipcode centroids, significant interactions between zipcode centroid distance and urban categorization and zipcode centroid distance and state of residence were observed in Models II ($p=0.0012$) and III ($p<0.0001$), respectively.

We applied Model II to Testing and stratified data for validation. The minimal R^2 value for Model II when applied to Testing data was 0.894 and was 0.779 when applied to stratified data (Table 5).

Comparing Accuracy of Prediction Equations

Prediction Models for Driving Distance

Predicting driving distance by straight line distance between residential and hospital

locations: Summarizing across Testing data, when distances were predicted using Model I, urban driving distances were overestimated by 14.1% (1.1 miles). Suburban distances were overestimated by 10.7% (1.7 miles). Rural driving distances were overestimated by 20.3% (3.4 miles). Predicted distances for urban residences were improved in Model II (overestimating 5.0%, 0.4 miles), when urban categorization was included in the prediction model. Suburban and rural predicted driving distances were less accurate, overestimated by 18.6% (3.0 miles) and 23.9% (4.0 miles), respectively.

Predicting driving distance by straight line distance between zipcode centroids:

Using distance between zipcode centroids alone, the predicted driving distance between residence and hospital in rural residences were overestimated by 19.0% (3.20 miles) compared to 39.9% (6.7 miles) for Model II. Suburban Model I predicted values were 23.1% (3.7 miles) larger than observed values while Model II predicted values were only 11.9% (1.9 miles) greater. Model I overestimated driving distances from urban locations by 27.2% (2.0 miles) while Model II overestimated distances by 11.3% (0.8 miles; Table 6).

Prediction Models for Driving Time

Predicting driving time by straight line distance between residential and hospital locations:

Predicted times from Model I overestimated for urban residence times 31.1% (3.8 minutes), suburban time by 25.1% (5.3 minutes), and rural residence times by 38.1% (9.8 minutes). Model II improved predicted values for urban (overestimated 24.6%, 3.0 minutes) and suburban (overestimated by 19.5%, 4.2 minutes) driving times. Rural times were overestimated by 52.1% (13.4 minutes) in Model II.

Predicting driving time by straight line distance between zipcode centroids: In rural areas, Model I predicted values overestimated driving time by 36.5% (9.4 minutes) while Model II overestimated time by 58.2% (15.0 minutes). In suburban areas, Model II had more accurate predicted values, overestimating driving time by an average of 17.6% (3.8 minutes) while Model I had an estimated 31.8% (6.8 minutes) overestimation. In urban areas, Model I overestimated time by 38.5% (4.7 minutes) and Model II overestimated driving time by 23.0% (2.8 minutes; Table 6).

Discussion

In a sample of Medicare beneficiaries living in Illinois, North Carolina, Ohio, or Tennessee at the time of elective TKR, we predicted the MapQuest driving distance and driving time between subject residence and hospital of TKR surgery using functions of straight line distances between residential and hospital locations, straight line distances between zipcode centroids, urban categorization of residence, and state of residence. Based on model AIC and R^2 values, four “best” models, two for each outcome, were selected:

- 1) Driving distance predicted by straight line distance between residential and hospital locations
- 2) Driving distance predicted by straight line zipcode centroid distance and urban categorization of residence
- 3) Driving time predicted by straight line distance between residential and hospital locations, urban categorization of residence, interaction between distance and urban categorization
- 4) Driving time predicted by straight line zipcode centroid distance, urban categorization of residence, interaction between zipcode distance and urban categorization

In models applied to Testing data and stratified subsamples, unadjusted R^2 values for driving distance and time were at least 0.911 and 0.822, respectively, when predicted by functions of residential/hospital distance, and 0.850 and 0.779 when predicted by functions of zipcode centroid distance.

While the R^2 values from the model validation indicated that the selected models fit the Testing data well, we observed that, on average, the models overestimated driving distance and driving time. Comparing relative bias, a model predicting driving distance from distance between residential and hospital locations alone had lower relative bias for suburban and rural areas than a model also including urban categorization of residence. In contrast, the addition of urban categorization improved models predicting driving distance from zipcode centroid distances. For both distance between residence/hospital locations and between zipcode centroids, the overestimation of driving distances in rural areas was increased when urban categorization was included in the models.

When predicting driving time, the addition of urban categorization substantially improved the accuracy of predicted urban and suburban residence driving times. Simpler models, including distance between residential and hospital location only or zipcode centroid distance only, outperformed more complex models in rural areas with smaller relative biases.

In general, subjects living in rural residences had longer driving distances and times than urban or suburban residences. For all three urban categories the distributions of observed driving distance and driving time were right-skewed with long tails. As a result, the mean values of distances and times for the three groups were larger than the median values with the largest differences observed for rural residences. When urban categorization was included in the regression equations, the rural residence intercept was greater than the intercept in models excluding urban categorization. Rural subjects who attended hospitals that were a short distance away from their homes had driving distances that were smaller than the inflated rural residence intercepts causing an increased overestimation of rural distances in models adjusting for urban categorization. An implication for future research is that if investigators are primarily interested in estimating distances in rural settings, these equations may not be appropriate. Estimation equations based solely on rural distances and equations using median distances in place of the ordinary least squares methods that rely on mean distances should be examined.

This study had a few limitations, the first of which being that we assumed that subjects drove to the hospital. One subject (in Testing 1) lived in North Carolina and traveled over 1,000 miles to Illinois for treatment; however the actual mode of transportation is unknown. It is possible that subjects with longer distances to care received services while at part-time or vacation residences, far from their listed billing addresses.

Our study was limited to Medicare recipients at least 65 years old who received a TKR while residing in one of four U.S. states. We selected Medicare recipients because Medicare claims data include precise residential address and the selected states had low proportions of Medicare HMO, making the reporting of claims mandatory. Total knee replacement was selected as the procedure of interest because it is a common procedure among Medicare recipients and distance to care is a concern for clinicians and policy makers as hospital choice impacts the probability of needing a second surgery (20). The four states also provided a diverse mix of urban, suburban, and rural population distributions from states located in either Midwestern (OH, IL) or Southern (NC, TN) U.S. Census regions. We observed consistent results when prediction equations were applied across stratified subsamples. The evaluation of this estimation method in a national cohort, with a more diverse sample of population densities, and subjects across the age continuum is left for future research.

There are several implications from this research. Precise residential and hospital locations may be available from medical records or Medicare claims data. After computing straight line distance between residential and hospital locations, researchers can apply the prediction models proposed in this study to estimate driving distance and driving time with reasonable accuracy. The addition of urban categorization, available from U.S. Census data, improves estimates for driving time but is unnecessary when estimating driving distance.

Distance measures based on zipcode centroids are known to overweight locations near boundaries as residences and hospitals located in different zipcodes may actually be very close to one another, resulting in estimates that are longer than the true distances (1). Despite this, when examining access to health care, researchers often use zipcode centroid distances as an approximation of driving distance or driving time between precise geographic

locations (3–6). For urban and suburban areas the zipcode centroids may be near enough to subject residences that they provide adequate spatial resolution for such estimates. For rural areas, however, the distances between residence and zipcode centroids are larger as rural zipcodes have larger square mileage than urban or suburban zipcodes, providing coarse spatial resolution. Accounting for urban categorization of residential census tract improved the prediction model estimates of driving distance and driving time greatly for urban and suburban data. While similar results were not observed for rural areas, the improvement of urban and suburban estimates provide a better estimation of driving distance and time than is currently available.

In general, aggregate measures of distance such as distances between zipcode centroids provide adequate proxies for more detailed measures of driving distance and driving time. When supplemented by U.S. Census information, such as urban categorization of census tracts, the estimates are enhanced, particularly for suburban and urban locations. The importance of urban categorization in relation to distances between zipcode centroids is not surprising. Zipcodes boundaries are defined by functions of square mileage and population density. Distances between neighboring zipcode centroids depends on the urban categorization of the areas and, therefore, it should be accounted for when producing distance and time estimates based on zipcode centroid distances. In health policy research, the addition of census tract urban categorization to distance-based studies will improve estimates and provide better intuition regarding analyses of spatial accessibility to health care.

Acknowledgments

Funding Agency: Grants 5T32AR055885-03, K24AR057827, P60AR47782 from the National Institute of Arthritis and Musculoskeletal and Skin Diseases, National Institute of Health

References

1. Guagliardo MF. Spatial accessibility of primary care: concepts, methods and challenges. *International Journal of Health Geographics*. 2004;3. [PubMed: 14987337]
2. Al-Taiar A, Clark A, Longenecker JC, et al. Physical accessibility and utilization of health services in Yemen. *International Journal of Health Geographics*. 2010;9. [PubMed: 20156361]
3. Goodman DC, Fisher E, Stukel TA, et al. The Distance to Community Medical Care and the Likelihood of Hospitalization: Is Closer Always Better? *American Journal of Public Health*. 1997;87.
4. Jordan H, Roderick P, Martin D, et al. Distance, rurality and the need for care: access to health services in South West England. *International Journal of Health Geographics*. 2004;3. [PubMed: 14987337]
5. Mayer ML. Are We There Yet? Distance to Care and Relative Supply Among Pediatric Medical Subspecialties. *Pediatrics*. 2006;118. [PubMed: 16510555]
6. Piette JD, Moos RH. The Influence of Distance on Ambulatory Care Use, Death, and Readmission Following a Myocardial Infarction. *Health Services Research*. 1996;31.
7. Nicholl J, West J, Goodacre S, et al. The relationship between distance to hospital and patient mortality in emergencies: an observational study. *Emergency Medical Journal*. 2004; 24:665–668.
8. Olson K, Grannis S, Mandi K. Privacy Protection Versus Cluster Detection in Spatial Epidemiology. *American Journal of Public Health*. 2006; 96:2002–2008. [PubMed: 17018828]
9. Ozonoff A, Jeffery C, Manjourides J, et al. Effect of spatial resolution on cluster detection: a simulation study. *International Journal of Health Geographics*. 2007;6. [PubMed: 17341310]
10. Webster T, Vieira V, Weinberg J, et al. Method for mapping population-based case-control studies: an application using generalized additive models. *International Journal of Health Geographics*. 2006;5. [PubMed: 16420702]

11. Morgenstern, H. Ecologic Study. In: Armitage, P.; Colton, T., editors. Encyclopedia of Biostatistics. John Wiley & Sons; Chichester, England: 2005. p. 1567-1588.
12. Elliott, P.; Wakefield, JC.; Best, NG., et al. Spatial epidemiology: methods and applications. In: Elliott, P.; Wakefield, J.; Best, N., et al., editors. Spatial Epidemiology: Methods and Applications. Oxford University Press; New York: 2000.
13. Boulos MNK, Curtis AJ, AbdelMalik P. Musings on privacy issues in health research involving disaggregate geographic data about individuals. International Journal of Health Geographics. 2009;8. [PubMed: 19208225]
14. Losina E, Plerhoples T, Fossel AH, et al. Offering Patients the Opportunity to Choose Their Hospital For Total Knee Replacement: Impact on Satisfaction With the Surgery. Arthritis & Rheumatism. 2005; 53:646–652. [PubMed: 16208651]
15. Inc M. MapQuest Developer Network. 2010. Available at: <http://developer.mapquest.com/2010>
16. Bellesfield, KJ.; Campbell, TL. Methods and Apparatus for Displaying a Travel Route and/or Generating a List of Places of Interest Located Near the Travel Route. United States Patent and Trademark Office; MapQuest.com, Inc.; 2002.
17. MapQuest Inc.. [Accessed November, 2010] MapQuest Developer Network. 2010. Available at: <http://developer.mapquest.com/>.
18. SAS for Windows [computer program]. Version 9.2. SAS Institute, Inc.; Cary, NC: 2008.
19. Weisberg, S. Applied Linear Regression. John Wiley & Sons, Inc.; Hoboken, New Jersey: 2005.
20. Katz JN, Barrett J, Mahomed NN, et al. Association between hospital and surgeon procedure volume and the outcomes of total knee replacement. J Bone Joint Surg Am. 2004; 86-A:1909–1919. [PubMed: 15342752]

Table 1

Residential distributions of total sample, Training, and Testing data by state and urban residence

	Total N(%)	Training N(%)	Testing 1 N(%)	Testing 2 N(%)	Testing 3 N(%)
Total	1135	284	286	273	292
Residence					
Urban	314(27.7)	88(31.1)	85(29.7)	68(24.9)	73(25.0)
Suburban	243(21.4)	63(22.3)	54(18.9)	56(20.5)	70(24.0)
Rural	577(50.8)	132(46.5)	147(51.4)	149(54.6)	149(51.0)
Illinois					
Residence					
Urban	364(32.1)	96(33.8)	98(34.3)	80(29.3)	90(30.8)
Suburban	182(50.1)	48(50.5)	53(54.1)	43(53.8)	38(42.2)
Rural	76(20.9)	17(17.9)	15(15.3)	18(22.5)	26(28.9)
	105(28.9)	30(31.3)	30(30.6)	19(23.8)	26(28.9)
North Carolina					
Residence					
Urban	204(18.0)	48(16.9)	55(19.2)	47(17.2)	54(18.5)
Suburban	15(7.4)	6(12.5)	2(3.6)	1(2.1)	6(11.1)
Rural	30(14.7)	7(14.6)	8(14.6)	3(6.4)	12(22.2)
	159(77.9)	35(72.9)	45(81.8)	43(91.5)	36(66.7)
Ohio					
Residence					
Urban	387(34.1)	98(34.5)	85(29.7)	102(37.4)	102(34.9)
Suburban	92(23.8)	26(26.5)	22(25.9)	20(19.6)	24(23.5)
Rural	99(25.6)	28(28.6)	20(23.5)	27(26.5)	24(23.5)
	196(50.7)	44(44.9)	43(50.6)	55(53.9)	54(52.9)
Tennessee					
Residence					
Urban	180(15.9)	42(14.8)	48(16.8)	44(16.1)	46(15.8)
Suburban	25(13.9)	8(19.1)	8(16.7)	4(9.1)	5(10.9)
Rural	38(21.1)	11(26.2)	11(22.9)	8(18.2)	8(17.4)
	117(65.0)	23(54.8)	29(60.4)	32(72.7)	33(71.7)

Table 2

Models predicting MapQuest driving distance and driving time from straight line distance between residential/hospital address locations (precise point distance) and straight line distance between zipcode centroids

	MapQuest driving distance by distance between residential/hospital locations	MapQuest driving distance by zipcode centroid distance	MapQuest driving time by distance between residential/hospital locations	MapQuest driving time by zipcode centroid distance
Model I	RHDistance [*]	ZipDistance [^]	RHDistance [*]	ZipDistance [^]
Model II	RHDistance + Urban Category + Possible interaction [†]	ZipDistance + Urban Category + Possible interaction [†]	RHDistance + Urban Category + Possible interaction [†]	ZipDistance + Urban Category + Possible interaction [†]
Model III	RHDistance + Urban Category + Stata + Possible interactions ^{†‡}	ZipDistance + Urban Category + Stata + Possible interactions ^{†‡}	RHDistance + Urban Category + Stata + Possible interactions ^{†‡}	ZipDistance + Urban Category + Stata + Possible interactions ^{†‡}

* RHDistance = straight line distance between precise residential and hospital addresses

[^] ZipDistance = straight line distance between zipcode centroids

[†] Hypothesis test performed to evaluate possible interaction between distance measure and urban categorization; if $p < 0.05$ then interaction is included in model

[‡] Hypothesis test performed to evaluate possible interaction between distance measure and state of residence; if $p < 0.05$ then interaction is included in model

Table 3

Summary statistics and Pearson's correlation coefficients (r)[†] of distance measures in training data by state

	Mean $\pm$ SD (Median)	Zipcode centroid distance Pearson's r	MapQuest driving distance Pearson's r	MapQuest driving time Pearson's r
Total				
Straight line distance between residential/hospital locations	10.3 $\pm$ 14.4(5.0)	0.989	0.986	0.951
Zipcode centroid straight line distance	10.0 $\pm$ 14.7(5.3)	---	0.981	0.945
MapQuest driving distance	13.4 $\pm$ 17.9(6.8)	---	---	0.970
MapQuest driving time	20.6 $\pm$ 22.7(12.0)	---	---	---

[†] All correlations significant at the $\alpha = 0.001$ level.

between distance between residential/

MapQuest driving time by distance between residential/hospital locations	MapQuest driving time by distance between zipcode centroids
$5.15+(1.50)*RHDistance$ 0.904 1911.1	$6.03+(1.46)*ZipDistance$ 0.894 1941.8
$5.60+(1.96)*Rural+(-0.63)*Suburban+(1.22)*RHDistance+(0.32)*Rural*RHDistance+(0.36)*Suburban*RHDistance$ 0.914 1887.0	$5.60+(1.96)*Rural+(-0.63)*Suburban+(1.22)*ZipDistance+(0.28)*Rural*ZipDistance+(0.22)*Suburban*ZipDistance$ 0.904 1917.4
$4.279+(1.25)*Rural+(-1.15)*Suburban+(1.38)*RHDistance+(0.38)*Rural*RHDistance+(0.42)*Suburban*RHDistance$ $5.67+(1.25)*Rural+(-1.15)*Suburban+(1.26)*RHDistance+(0.38)*Rural*RHDistance+(0.42)*Suburban*RHDistance$ $7.22+(1.25)*Rural+(-1.15)*Suburban+(0.98)*RHDistance+(0.38)*Rural*RHDistance+(0.42)*Suburban*RHDistance$ $7.35+(1.25)*Rural+(-1.15)*Suburban+(1.11)*RHDistance+(0.38)*Rural*RHDistance+(0.42)*Suburban*RHDistance$ 0.921 1866.5	$4.279+(1.25)*Rural+(-1.15)*Suburban+(1.38)*ZipDistance+(0.31)*Rural*ZipDistance+(0.25)*Suburban*ZipDistance$ $5.67+(1.25)*Rural+(-1.15)*Suburban+(1.26)*ZipDistance+(0.31)*Rural*ZipDistance+(0.25)*Suburban*ZipDistance$ $7.22+(1.25)*Rural+(-1.15)*Suburban+(1.01)*ZipDistance+(0.31)*Rural*ZipDistance+(0.25)*Suburban*ZipDistance$ $7.35+(1.25)*Rural+(-1.15)*Suburban+(1.11)*ZipDistance+(0.31)*Rural*ZipDistance+(0.25)*Suburban*ZipDistance$ 0.911 1899.3

Table 5

Model R² values when applied to Testing and stratified data

MapQuest Driving distance	Model I including straight line distance between residential/hospital location				Model I including straight line zipcode centroid distance							
	Testing 1		Testing 2		Testing 3		Testing 1		Testing 2		Testing 3	
	N	R ²	N	R ²	N	R ²	N	R ²	N	R ²	N	R ²
Total sample	286	0.998	273	0.973	292	0.981	286	0.996	273	0.948	292	0.967
Residence												
Urban	85	0.958	68	0.972	73	0.911	85	0.927	68	0.900	73	0.850
Suburban	54	>0.999	56	0.973	70	0.967	54	0.999	56	0.909	70	0.955
Rural	147	0.972	149	0.971	149	0.987	147	0.944	149	0.947	149	0.974
State												
Illinois	98	0.982	80	0.967	90	0.988	98	0.970	80	0.931	90	0.982
North Carolina	55	>0.999	47	0.950	54	0.946	55	0.999	47	0.891	54	0.895
Ohio	85	0.983	102	0.981	102	0.982	85	0.968	102	0.968	102	0.968
Tennessee	48	0.963	44	0.972	46	0.992	48	0.945	44	0.938	46	0.982

MapQuest Driving time	Model II straight line distance between residential/hospital location and urban categorization of residential census tract				Model II including straight line zipcode centroid distance and urban categorization of residential census tract							
	N	R ²	N	R ²	N	R ²	N	R ²	N	R ²	N	R ²
Total sample	286	0.989	273	0.915	292	0.948	286	0.988	273	0.894	292	0.939
Residence												
Urban	85	0.868	68	0.891	73	0.863	85	0.838	68	0.825	73	0.803
Suburban	54	0.999	56	0.888	70	0.947	54	0.999	56	0.828	70	0.933
Rural	147	0.894	149	0.908	149	0.953	147	0.864	149	0.886	149	0.946
State												
Illinois	98	0.927	80	0.881	90	0.968	98	0.920	80	0.847	90	0.963
North Carolina	55	0.998	47	0.822	54	0.894	55	0.997	47	0.779	54	0.848
Ohio	85	0.926	102	0.934	102	0.942	85	0.907	102	0.929	102	0.940

MapQuest Driving time	Model II straight line distance between residential/hospital location and urban categorization of residential census tract						Model II including straight line zipcode centroid distance and urban categorization of residential census tract					
	N	R ²	N	R ²	N	R ²	N	R ²	N	R ²	N	R ²
Tennessee	48	0.879	44	0.937	46	0.978	48	0.867	44	0.889	46	0.979

Table 6

Relative bias of estimated driving distance and driving time from models I* and II^, estimates based on straight line distance between residential/hospital location and straight line distance between zipcode centroids

Distance	All Testing Data Mean Relative Bias [†] (Average # Miles/Minutes)		Testing 1 Mean Relative Bias [†] (Average # Miles/Minutes)		Testing 2 Mean Relative Bias [†] (Average # Miles/Minutes)		Testing 3 Mean Relative Bias [†] (Average # Miles/Minutes)	
	Precise Points	Zipcode	Precise Points	Zipcode	Precise Points	Zipcode	Precise Points	Zipcode
Urban								
Model I	-14.1% (-1.1)	-27.2% (-2.0)	-14.9% (-1.1)	-29.9% (-2.3)	-15.5% (-1.0)	-32.4% (-2.2)	-11.9% (-1.0)	-19.3% (-1.6)
Model II	-5.0% (-0.4)	-11.3% (-0.8)	-5.4% (-0.4)	-13.1% (-1.0)	-5.7% (-0.4)	-15.1% (-1.0)	-4.0% (-0.3)	-5.5% (-0.4)
Suburban								
Model I	-10.7% (-1.7)	-23.1% (-3.7)	-7.3% (-2.1)	-24.2% (-7.8)	-15.6% (-1.2)	-25.5% (-2.0)	-9.4% (-1.2)	-17.8% (-2.3)
Model II	-18.6% (-3.0)	-11.9% (-1.9)	-14.4% (-4.1)	-27.3% (-4.9)	-24.5% (-1.9)	-13.1% (-1.0)	-17.0% (-2.2)	-7.0% (-0.9)
Rural								
Model I	-20.3% (-3.4)	-19.0% (-3.2)	-19.0% (-3.2)	-6.5% (-1.1)	-24.3% (-4.1)	-27.7% (-4.6)	-17.6% (-3.0)	-22.6% (-3.9)
Model II	-23.9% (-4.0)	-39.9% (-6.7)	-22.0% (-3.7)	-24.2% (-4.1)	-28.6% (-4.8)	-24.2% (-8.7)	-21.1% (-3.6)	-43.2% (-7.4)

Time	Precise Points		Zipcode		Precise Points		Zipcode	
	Precise Points	Zipcode	Precise Points	Zipcode	Precise Points	Zipcode	Precise Points	Zipcode
Urban								
Model I	-31.1% (-3.8)	-38.5% (-4.7)	-35.5% (-4.4)	-43.7% (-5.4)	-32.8% (-3.7)	-43.0% (-4.9)	-24.5% (-3.2)	-28.2% (-3.7)
Model II	-24.6% (-3.0)	-23.0% (-2.8)	-29.5% (-3.6)	-27.9% (-3.4)	-27.0% (-3.1)	-27.1% (-3.1)	-16.8% (-2.2)	-13.4% (-1.7)
Suburban								
Model I	-25.1% (-5.3)	-31.8% (-6.8)	-20.6% (-7.0)	-30.8% (-10.4)	-28.9% (-3.7)	-34.9% (-4.5)	-25.4% (-4.7)	-30.2% (-5.6)
Model II	-19.5% (-4.2)	-17.6% (-3.8)	-15.5% (-5.2)	-17.2% (-5.8)	-22.4% (-2.9)	-19.6% (-2.5)	-20.3% (-3.8)	-16.4% (-3.0)
Rural								
Model I	-38.1% (-9.8)	-36.5% (-9.4)	-30.7% (-7.9)	-22.3% (-5.7)	-43.5% (-11.5)	-43.9% (-11.6)	-40.1% (-10.1)	-43.2% (-10.9)
Model II	-52.1% (-13.4)	-58.2% (-15.0)	-43.0% (-11.0)	-41.0% (-10.5)	-58.7% (-15.5)	-67.8% (-17.9)	-54.5% (-13.7)	-65.7% (-16.5)

* Model I estimates from model including straight line or zipcode centroid distances only

^ Model II estimates from model including straight line or zipcode centroid distances and urban categorization of residence

$$\text{Relative bias} = 100 \frac{(\text{observed} - \text{estimated})}{\text{observed}} \%$$