Journal of Urban Health: Bulletin of the New York Academy of Medicine, Vol. 87, No. 4
doi:10.1007/s11524-010-9458-0

Concordance of Commercial Data Sources for Neighborhood-Effects Studies

Christine M. Hoehner and Mario Schootman

ABSTRACT *Growing evidence supports a relationship between neighborhood-level characteristics and important health outcomes. One source of neighborhood data includes commercial databases integrated with geographic information systems to measure availability of certain types of businesses or destinations that may have either favorable or adverse effects on health outcomes; however, the quality of these data sources is generally unknown. This study assessed the concordance of two commercial databases for ascertaining the presence, locations, and characteristics of businesses. Businesses in the St. Louis, Missouri area were selected based on their four-digit Standard Industrial Classification (SIC) codes and classified into 14 business categories. Business listings in the two commercial databases were matched by standardized business name within specified distances. Concordance and coverage measures were calculated using capture–recapture methods for all businesses and by business type, with further stratification by census-tract-level population density, percent below poverty, and racial composition. For matched listings, distance between listings and agreement in four-digit SIC code, sales volume, and employee size were calculated. Overall, the percent agreement was 32% between the databases. Concordance and coverage estimates were lowest for health-care facilities and leisure/entertainment businesses; highest for popular walking destinations, eating places, and alcohol/tobacco establishments; and varied somewhat by population density. The mean distance (SD) between matched listings was 108.2 (179.0) m with varying levels of agreement in four-digit SIC (percent agreement=84.6%), employee size (weighted kappa=0.63), and sales volume (weighted kappa=0.04). Researchers should cautiously interpret findings when using these commercial databases to yield measures of the neighborhood environment.*

KEYWORDS *Concordance, Geographic Information Systems, Capture–Recapture Methods, Neighborhood Studies, Built Environment*

INTRODUCTION

Growing evidence supports a relationship between neighborhood-level characteristics and important health outcomes.[1] One secondary source of neighborhood data includes commercial and administrative databases integrated with geographic information systems to measure availability of certain types of businesses or destinations that may have either favorable (e.g., grocery stores) or adverse (e.g., liquor outlets) effects on health outcomes. For example, locations of food stores and restaurants provide information on food choices and access to healthy foods, which may be particularly important for low-income and minority populations.[2–4]

Hoehner is with the Department of Surgery and the Alvin J. Siteman Cancer Center, School of Medicine, Washington University, St. Louis, MO, USA; Schootman is with the Departments of Medicine and Pediatrics and Alvin J. Siteman Cancer Center, School of Medicine, Washington University, St. Louis, MO, USA.

Correspondence: Christine M. Hoehner, Department of Surgery and the Alvin J. Siteman Cancer Center, School of Medicine, Washington University, St. Louis, MO, USA. (E-mail: hoehnerc@wudosis.wustl.edu)

Availability of places that provide opportunities for physical activity have been associated with individual physical activity and obesity levels.[5–8] Building on this work, the neighborhood service environment, as measured by availability of various types of businesses and organizations hypothesized as favorable or unfavorable to health, has been examined for its association with health outcomes.[9,10]

Researchers are beginning to assess the quality of secondary data sources with business information,[11–14] given that these data sources are increasingly being used in research with some expressing concerns about the accuracy and completeness of these data sources.[5,15–18] Information on the completeness and accuracy of multiple data sources for ascertaining local businesses is important not only to health researchers, but also to community organizations, policy-makers, transportation planners, and others who seek to map local businesses and resources.

The aim of this study was to assess the concordance of two commercial databases for ascertaining the presence, locations, and characteristics of businesses by type and area-level poverty status, racial composition, and population density. Capture–recapture methods were applied to estimate the total number of business listings, including listings excluded from either database. These methods are used to estimate a population size when a census may be infeasible or impossible to conduct.[19–21] Originally used for population estimation and for wildlife research and management, this technique has been applied to epidemiology for estimating incidence and prevalence of various diseases and health-related problems using data collected by two or more incomplete sources (e.g., hospital records and death certificates).[19–22] For the purposes of the present study, the capture–recapture method allowed estimation of the coverage (i.e., sensitivity or extent to which the databases included the complete number of listings) of two databases for ascertaining businesses.

METHODS

Study Area

The study area consisted of the City of St. Louis and St. Louis County, Missouri, United States. This area includes 590 square miles, 286 census tracts, and 1,686,724 people.[23]

Data Sources

Business names and locations were obtained from two of the major commercial vendors of business databases in the United States.

- *Database A* includes data from the InfoUSA database bundled with the ArcGIS 9.2.5 Business Analyst software (ESRI, Redlands, California, USA). InfoUSA data are compiled from phone books, business directories, 10Ks and Securities and Exchange Commission information, government data, business magazines, newsletters and newspapers, and information from the U.S. Postal Service, verified by calling businesses.[24] The InfoUSA database includes business name, industry description (i.e., Standard Industry Classification [SIC] or North American Industry Classification System [NAICS]), sales, employees, and location (latitude and longitude) based on geocoding to Tele Atlas address and street databases.[24] The addresses in the database bundled with the Business Analyst software are pre-geocoded. Latitude and longitude are provided, but addresses are removed. The data were current as of January 2008.
- *Database B* includes data purchased from Dun and Bradstreet, a business information provider. Companies apply for free for credit purposes. According

to the company's website, the data are collected, aggregated, edited, and verified "from thousands of sources daily."[25] Data include address, a primary four-digit SIC code, a primary six-digit NAICS code, company names, business descriptions, number of employees, sales volume, and square footage of buildings. The database for this study included businesses active through the year 2007, with a cost of approximately $4,800.

Businesses were selected for inclusion in this study based on their four-digit SIC codes and classified as in previous studies[7,9,26,27] into five broad and nine more specific categories (see Table 1 for categories and Appendix for SIC codes).

Geocoding Addresses

Records from the database with addresses (database B) were geocoded using the 2005 Streetmap Extension of ArcGIS. Any records not coded or coded with scores less than 80 (n=2,646, 13.6%) were geocoded again using the TeleAtlas web-based geocoder.

Matching Businesses

Because business addresses were not included in database A, business listings in the two databases were matched by standardized business name within specified distances. Initially, business matches were sought within nine adjacent 1,000-m cells surrounding a database B business location (hereafter, 1,000-m grids; Figure 1). When this system produced multiple matches for a single business listing, a tiered approach was adopted, whereby matches for business listings were sought within 10-m grids (tier 1), followed by 100-m grids (tier 2) and then 1,000-m grids (tier 3). Duplicates within a single database and matches between databases were removed prior to the initiation of subsequent tiers.

Multiple methods were employed to standardize the business name to identify matches within the 10-, 100-, and 1,000-m grids, including the following:

(1) *Finding common names for chain businesses in multiple locations*. Business names were parsed into words in sequence. Commonly occurring single-word names and two- to five-word combinations were identified to select standardized names. Most multi-location businesses (~4,500) were matched in this manner.
(2) *Reordering business names for listings that included an individual's name as the business name or part of the name*. One database would consistently use the reverse order from the other source (last name first vs. first name first), so to facilitate matching, the standard name for one source was reordered.
(3) *Searching similar character strings*. The COMPLEV function in the SAS software program was used to calculate the Levenshtein distance between two strings—a mathematical formula for the similarity of character strings. This function identified the top ten most likely matches for unmatched businesses within the tier 3 grids.

Analysis

Concordance and coverage estimates were calculated for all businesses and by business type, with further stratification by census-tract-level population density, percent below poverty, and race based on cutpoints used by others (Table 1).[28,29]

TABLE 1 Agreement and coverage[a] of two commercial databases, by business type and area-level characteristics

Classifications by business type and area-level[b] characteristics	Database A	Database B	Percent difference	Percent agreement	Coverage of database A (95% CI)	Coverage of database B (95% CI)	Coverage of both databases (95% CI)
All businesses	18,199	16,615	8.7	32.0	50.8 (50.0–51.5)	46.3 (45.6–47.1)	73.6 (73.0–74.1)
Broad business categories							
Physical activity facilities	772	698	9.6	34.4	53.9 (50.4–57.4)	48.7 (45.0–52.4)	76.3 (73.8–78.8)
Popular walking destinations	4,416	3,798	14.0	46.0	68.2 (66.8–69.5)	58.6 (57.1–60.2)	86.8 (85.9–87.7)
Services with undesirable amenities	1,755	1,692	3.6	35.3	53.1 (50.8–55.5)	51.2 (48.9–53.6)	77.1 (75.5–78.8)
Services providing health care	6,255	5,308	15.1	24.5	42.9 (41.7–44.1)	36.4 (35.1–37.7)	63.7 (62.7–64.7)
Services promoting social engagement	9,196	8,984	2.3	38.0	55.8 (54.8–56.8)	54.5 (53.5–55.5)	79.9 (79.2–80.5)
Specific business categories							
Library and post offices	166	118	28.9	35.2	62.9 (55.5–70.2)	44.7 (35.7–53.7)	79.3 (73.8–84.8)
Food stores	866	740	14.5	38.2	60.0 (56.8–63.3)	51.3 (47.7–54.9)	80.5 (78.2–82.8)
Eating places	2,820	2,605	7.6	51.0	70.4 (68.7–72.1)	65.0 (63.2–66.8)	89.6 (88.6–90.6)
Alcohol and tobacco establishments	558	523	6.3	50.6	69.4 (65.6–73.3)	65.1 (61.0–69.2)	89.3 (87.0–91.6)
Other retail businesses	1,984	1,866	5.9	27.4	44.4 (42.3–46.6)	41.8 (39.6–44.0)	67.6 (66.0–69.3)
Banks	517	299	42.2	36.5	73.0 (69.1–76.8)	42.2 (36.6–47.8)	84.3 (81.4–87.2)
Beauty and barber shops	2,007	1,898	5.4	41.2	60.0 (57.9–62.2)	56.8 (54.5–59.0)	82.7 (81.3–84.1)
Leisure and entertainment businesses	497	491	1.2	21.8	36.1 (31.9–40.4)	35.7 (31.5–39.9)	58.7 (55.3–62.1)
Religious and membership organizations	2,502	2,834	−13.3	34.6	48.4 (46.4–50.3)	54.8 (53.0–56.6)	76.7 (75.3–78.0)
Population density							
Low (<3,200 persons/square mile)	8,753	7,866	10.1	30.1	48.8 (47.8–49.9)	43.9 (42.8–45.0)	71.3 (70.5–72.1)
Medium (3,200–6,400 persons/ square mile)	6,425	5,886	8.4	33.6	52.6 (51.3–53.8)	48.1 (46.9–49.4)	75.4 (74.5–76.3)
High (>6,400 persons/square mile)	3,021	2,863	5.2	34.2	52.4 (50.6–54.1)	49.6 (47.8–51.5)	76.0 (74.7–77.3)
Percentage below poverty							
<10%	11,724	10,600	9.6	31.6	50.6 (49.7–51.5)	45.8 (44.8–46.7)	73.2 (72.5–73.9)
10–19.9%	3,198	2,981	6.8	32.3	50.6 (48.9–52.4)	47.2 (45.4–49.0)	73.9 (72.7–75.2)
≥20%	3,240	3,003	7.3	32.9	51.4 (49.7–53.1)	47.7 (45.9–49.4)	74.6 (73.3–75.8)
Racial distribution							
>75% white	11,275	10,066	10.7	31.8	51.2 (50.3–52.1)	45.7 (44.7–46.7)	73.5 (72.8–74.2)
Mixed race	4,725	4,458	5.7	31.1	48.9 (47.5–50.4)	46.2 (44.7–47.6)	72.5 (71.5–73.5)
>75% black	2,162	2,060	4.7	34.5	52.6 (50.5–54.7)	50.1 (47.9–52.3)	76.3 (74.8–77.8)

[a]Derived using capture–recapture methods

[b]Derived from the U.S. Census 2000 for census tracts encompassing the business location

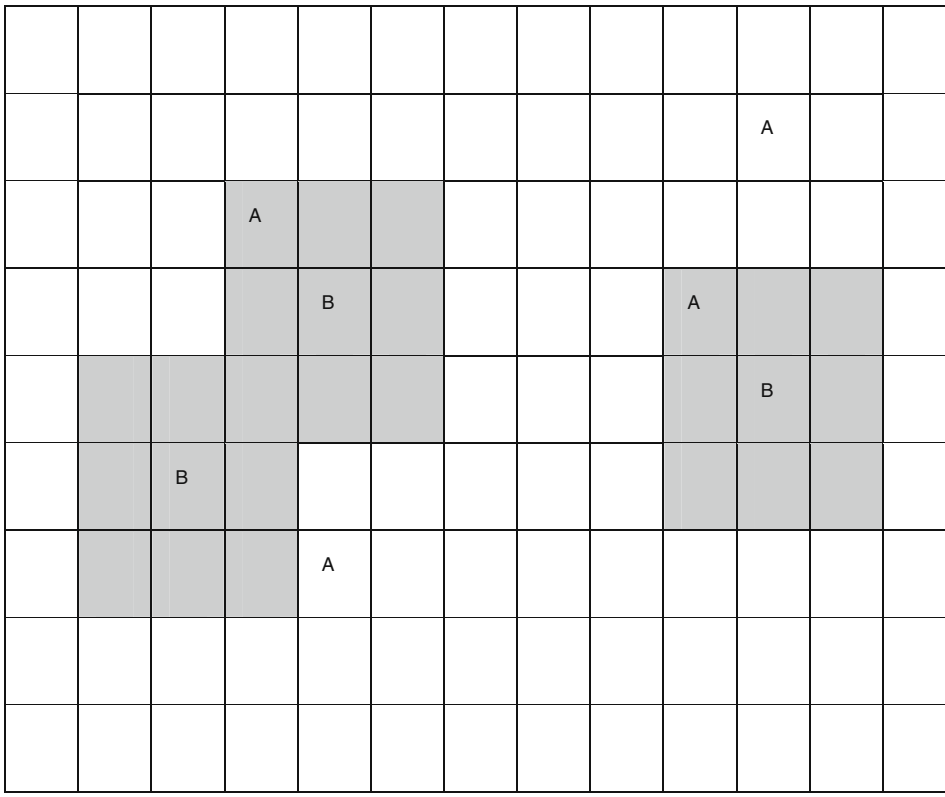

FIGURE 1. Illustration of grid system for identifying matches between business listings in two commercial databases. Note: The size of each individual cell was 10 m for tier 1, 100 m for tier 2, and 1,000 m for tier 3. The *shaded, nine-celled grids* illustrate the grids surrounding database B listings (denoted by the *letter B*). The *letter A* denotes listings from database A. The A listings *within the shaded area* are candidates for matching.

Capture–recapture methods were used to estimate the total number of listings (N) and the number of listings missed from both databases (x), based on the generic data structure in Figure 2.[19–22]

If we assume the databases are independent, then the probability of a business being present in database A if it is present in database B is equal to the probability of a business being present in database A if it is not present in database B:

$$P(\text{in A}|\text{inB}) = a/(a+b) = P(\text{in A}|\text{not in B}) = c/(c+x)$$

		Database B		
		Yes	No	Total
Database A	Yes	a	b	M
	No	c	x	
	Total	C		N

FIGURE 2. Generic data structure used to assess concordance and coverage of two commercial databases.

Rearranging the formula, we can calculate x:

$$\begin{aligned} a/(a+b) &= c/(c+x) \\ x &= (b \times c)/a \end{aligned} \tag{1}$$

Therefore,

$$N = a + b + c + x$$

or, using formula 1:

$$\begin{aligned} N &= a + b + c + (b \times c)/a \\ &= [(a+b) \times (a+c)]/a \end{aligned}$$

The following statistics were calculated:

1. % difference $= (M - C)/M \times 100\%$
2. % agreement $= a/(a+b+c) \times 100\%$
3. Coverage of Database A $= M/N \times 100\%$
4. Coverage of Database B $= C/N \times 100\%$
5. Coverage of both databases (i.e., % captured by either or both lists) $= (a+b+c)/N$

Concordance was also measured for attributes of the listings present in both databases (hereafter, matched listings). Distance between geocoded points was examined, as well as percent agreement for four-digit SIC code and weighted kappas for US Census-based categories of number of employees and sales volume (Table 2).

RESULTS

After excluding duplicates (189 in database A; 90 in database B), database A included 18,199 listings, and database B included 16,615 listings (8.7% difference; Table 1). With the exception of religious and membership organizations, database A contained more listings than database B across all business classifications.

The percent agreement between the databases was 32.0% for all types of businesses combined, ranging from 21.8% for leisure and entertainment businesses to 51.0% for eating places (Table 1). The coverage of database A exceeded the coverage of database B for all business categories except religious and membership organizations. Most of the differences in coverage estimates between databases were small, with the exception of those for libraries/post offices and banks. The coverage of both databases for all businesses combined was 73.6%, but ranged from 58.7% for leisure and entertainment businesses to 89.6% for eating places.

As shown in Table 1, agreement between databases for all businesses combined varied only slightly by area-level characteristics. The most apparent trend appeared for census-tract population density; agreement and coverage tended to increase with population density. A positive trend was also present, but much weaker, for the agreement and coverage estimates by poverty status.

For some business types, percent agreement, and coverage of the individual and combined databases varied by area-level characteristics (Figure 3). The most consistent trends were observed for census-tract population density, where the coverage of both databases differed by more than 10% between at least two strata.

TABLE 2 Distance in meters and agreement in four-digit SIC codes, business size and sales among business listings contained in both commercial databases[a]

	Distance in meters	Four-digit SIC code	No. of employees[b]	Sales volume[c]
Business type classification	Mean (SD)	Percent agreement	Weighted kappa (SD)	Weighted kappa (SD)
All businesses	108.2 (179.0)	84.6	0.63 (0.59–0.67)	0.04 (0.02–0.06)
Broad business categories				
Physical activity facilities	136.9 (198.4)	63.3	0.55 (0.43–0.67)	0.24 (0.11–0.36)
Popular walking destinations	103.0 (177.9)	87.6	0.66 (0.60–0.72)	0.02 (0.01–0.04)
Services with undesirable amenities	95.9 (173.8)	73.3	0.56 (0.48–0.64)	0.19 (0.11–0.27)
Services providing health care	127.5 (180.9)	76.3	0.55 (0.45–0.66)	0.03 (0.01–0.06)
Services promoting social engagement	102.1 (175.6)	84.5	0.63 (0.59–0.67)	0.03 (0–0.06)
Specific business categories				
Libraries and post offices	123.2 (202.9)	97.3	0.85 (0.71–1.00)	NA
Food stores	94.3 (177.4)	75.0	0.75 (0.61–0.88)	0.00 (−0.02–0.01)
Eating places	101.9 (174.7)	89.2	0.61 (0.53–0.70)	0.06 (0.01–0.11)
Alcohol and tobacco establishments	71.3 (117.7)	68.3	0.59 (0.49–0.68)	0.16 (0.06–0.26)
Other retail businesses	112.3 (189.5)	80.9	0.72 (0.65–0.79)	0.07 (0.01–0.13)
Banks	121.9 (206.2)	61.9	0.34 (0.14–0.55)	0.03 (−0.03–0.08)
Beauty and barber shops	89.6 (162.5)	88.9	0.56 (0.48–0.65)	0.13 (0.09–0.17)
Leisure and entertainment businesses	133.2 (240.8)	63.3	0.79 (0.57–1.00)	0.00 (−0.01–0.01)
Religious and membership organizations	97.2 (161.1)	82.3	0.50 (0.42–0.58)	0.03 (−0.01–0.07)

[a]N=8,434 matched listings. N=6,611 matched listings with nonmissing no. of employees (N=1,823 with missing no. of employees in database B); N=7,978 matched listings with nonmissing sales volume (N=456 with missing sales volume in database B)

[b]Categories for number of employees: 0, 1–4, 5–9, 10–19, 20–99, 100–499, 500+

[c]Categories for sales volume in thousands (K), millions (M), or billions (B): <$100K, $100–499K, $500–999K, $1–4.9M, $5–9.9M, $10–49.9M, $50–99.9M, $100–249.9M, $250–499.9M, $500–999.9M, $1–2.499B, ≥$2.5B

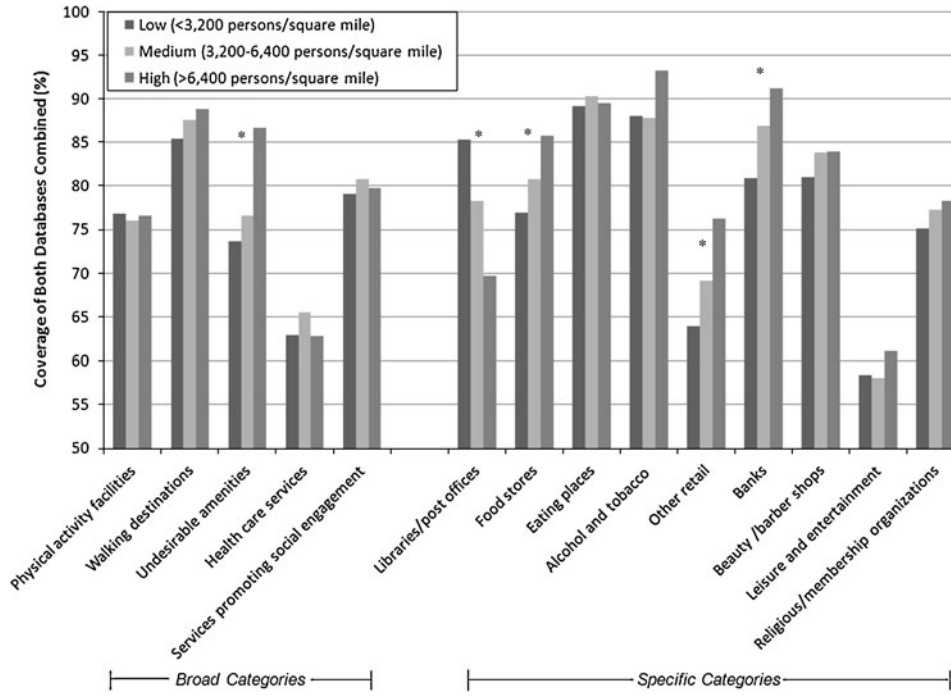

FIGURE 3. Coverage of both databases by business classification and population density of census tracts. *Asterisk* indicates >10% difference in coverage estimates between at least two strata.

Coverage estimates increased with increasing population density for businesses with undesirable amenities, banks, food stores, and retail businesses. An inverse association was observed for libraries and post offices. While some stratum-specific estimates for coverage of both databases varied by more than 10% for the poverty and race strata, the directions of the trends were inconsistent (data not shown). The only exceptions were observed for physical activity facilities where the coverage estimates varied by racial composition (77.3% for ≥75% white; 75.8% for mixed race; and 66.1% for ≥75% black) and for services with undesirable amenities where the coverage estimates increased with higher poverty levels (74.6% for <10% below poverty; 78.8% for 10–19.9% below poverty; and 83.4% for ≥20% below poverty). Similar patterns were observed for percent agreement.

Agreement in attributes of the 8,434 matched listings (distance, four-digit SIC code, number of employees, and sales volume) was also examined (Table 2). The mean (SD) distance between matched listings was 108.2 (179.0) m (~0.07 miles). The 95th percentile and maximum for distances between matched listings were 369.0 m (~0.23 miles) and 2,283.2 m (~1.4 miles), respectively (data not shown). The average distance between matched listings varied somewhat by business type, from 71.3 m difference for alcohol and tobacco establishments to 136.9 m difference for physical activity facilities.

The percent agreement for four-digit SIC was 84.6% for all businesses combined and ranged from 61.9% for banks to 97.3% for libraries and post offices. Agreement in number of employees was moderately high for all businesses combined (weighted kappa=0.63) with the weighted kappa for most business categories falling be 0.55 and 0.80. Agreement in sales volume was poor for all businesses combined (weighted kappa=0.04) and across all business categories (all

weighted kappas <0.25). Most of this discrepancy was attributed to differences in the distribution of matched listings with <$100K sales (27.6% of database A and 53.9% of database B).

DISCUSSION

Although commercial databases provide a feasible means to characterize availability and density of a variety of businesses, our study supports recent evidence[2,12–14] that data from commercial databases may contain substantial errors, with little agreement between one another. We found mostly fair agreement between databases (32%) which varied somewhat by business types and area-level characteristics. Agreement and coverage were highest for popular walking destinations, eating places, and alcohol and tobacco establishments. Perhaps these are businesses that benefit most from advertisement and registration in the databases. Agreement and coverage were lowest for health-care facilities and leisure and entertainment businesses. The low agreement among health-care facilities is likely attributable to the heterogeneity of businesses within this category, from individual health professionals to large hospitals.

Four published studies of smaller geographic areas (either selected census tracts or city blocks) assessed the quality of commercial data by comparing such data on physical activity facilities (e.g., health clubs, dance studios) and/or food stores (e.g., grocery stores, convenience stores) with field data[11–13] or governmental records.[14] These studies found moderate agreement for physical activity facilities[12,13] and moderate to high agreement for food stores.[11,13,14] We are aware of only two other studies that have directly compared databases A and B. The first compared aggregated counts of various classifications of food-related businesses in 50 zip codes in the Minneapolis/Saint Paul, Minnesota metropolitan area.[2] Although the counts for the broader businesses classifications for food and beverage stores and food services and drinking places showed only minor differences between databases A and B (–13% and –5%, respectively), large discrepancies in the counts of subtypes of these businesses were found. Another smaller study assessed concordance for retail services and personal care businesses in one zip code in west Miami Dade, Florida showing 46.5% agreement, with coverages of 84.2% for database A and 62.4% for database B.[30] Our findings extend these results by including many different types of businesses across a much larger geographic area.

Agreement and coverage of both databases varied slightly by census-tract population density, racial composition, and poverty level for some businesses. This finding differs from that observed by Boone et al. who found slightly higher agreement for physical activity facilities among nonurban vs. urban census tracts.[12] Paquet et al. observed no variation in agreement for food stores or physical activity facilities by area-level socioeconomic status.[13] Bader et al. found no consistent pattern between area-level sociodemographic characteristics and disagreement between field data and a commercial data source for the presence of various food outlets.[11] The reason for variation in agreement by population density in this study may be attributed to geocoding errors, particularly for addresses that were pre-geocoded in database A. Zip code centroids are often the default location when addresses cannot be located and may be a considerable distance from actual addresses, particularly in lower-density, suburban areas. Also, the low-to-moderate agreement in the number of

employees and sales volume for businesses present in both databases raises questions about the utility of these data for characterizing the size of businesses, as observed by others.[30]

Limitations of our study include possible outdated business listings, false negative matches, and false positive matches. We did not assess the extent of outdated business listings, but differential errors in ascertainment between the databases would result in an underestimation of the coverage of listings in the more accurate database. Despite the possibility of this error, coverage remained low for individual databases. Also, some matches may have been missed (false negatives) if their business name and locations differed between sources or fell outside the catchment area for detecting matches. Significant manual labor and programming skills were required for geocoding addresses (database B) and comparing business names at the varying catchment areas (approximately 500 hours of total person-time), so error is possible but unlikely to change the conclusions about the low levels of agreement. Field validation was infeasible given the number of businesses and our geographic area. This represents a significant limitation of this study because the direction of error for each database cannot be confirmed. False negative matches may have also resulted from businesses being differentially classified by four-digit SIC code between the data sources. Some discordant listings may be included in both databases but may have been classified by SIC codes excluded by our a priori selection in one of the databases.[12] Four-digit SIC codes are not standardized across databases.[2] Finally, different businesses may have been incorrectly matched (false positives) if they contained similar names or were branches of the same chain business within the catchment area. The fact that approximately 95% of matched listings were within 400 m of one another provided some support that the matches were true matches.

Despite these limitations, this study contributes important evidence about the low concordance of two of the most widely used commercial business databases. The geographic area and number of businesses examined exceed those of previous studies. Moreover, this study applied capture–recapture methods to estimate the total number and coverage of the databases—a promising method to estimate exposure to specific destinations when multiple, incomplete data sources are available.

To measure exposure to certain businesses for neighborhood-effects studies, researchers must select between existing databases or collect field data when feasible. Based on our findings, combining commercial databases may be impractical for studies covering large geographic areas when the databases lack standardized business classifications and common identifiers to match businesses between databases. Overall, health researchers should cautiously interpret findings when using either of these commercial databases to yield measures of the neighborhood service environment. Differences in agreement raise questions about differential misclassification when using these databases to characterize neighborhoods and their effects on health outcomes.

ACKNOWLEDGMENTS

We thank the Alvin J. Siteman Cancer Center at Barnes-Jewish Hospital and Washington University School of Medicine in St. Louis, Missouri, USA for the use of

the Health Behavior and Outreach Core. This study was supported in part by an American Cancer Society Mentored Research Scholar Grant (MRSG-07-016-01-CPPB) and by the National Cancer Institute (CA112159).

APPENDIX

APPENDIX Categorization of businesses by four-digit standard industrial classification (SIC) codes

Business type classification	Four-digit SIC codes
Broad business categories	
Physical activity facilities	7911, 7991, 7992, 7997, 7999
Popular walking destinations	4311, 5411, 5461, 5499, 5812, 5912, 6021, 6022, 6035, 6036, 6061
Services with undesirable amenities	5091, 5813, 5921, 5932, 5941, 5944, 5993, 7299
Services providing health care	5048, 5912, 5995, 5999, 7352, 7629, 8011, 8021, 8031, 8041, 8042, 8043, 8049, 8051, 8052, 8059, 8062, 8063, 8069, 8082, 8092, 8093, 8099
Services promoting social engagement	5812, 7231, 7241, 7829, 7832, 7911, 7922, 7929, 7933, 7941, 7948, 7991, 7992, 7993, 7996, 7997, 7999, 8231, 8322, 8399, 8412, 8422, 8611, 8621, 8631, 8641, 8651, 8661, 8699, 9441
Specific business categories	
Libraries and post offices	4311, 8231
Food stores	5411, 5461, 5499
Eating places	5812
Alcohol and tobacco establishments	5813, 5921, 5993
Other retail businesses	5091, 5912, 5932, 5941, 5944, 5995, 5999, 7629
Banks	6021, 6022, 6035, 6036, 6061
Beauty and barber shops	7231, 7241, 7299
Leisure and entertainment businesses	7829, 7832, 7922, 7929, 7933, 7941, 7948, 7993, 7996, 8412, 8422
Religious and membership organizations	8611, 8621, 8631, 8641, 8651, 8661, 8699

REFERENCES

1. Kawachi I, Berkman LF, eds. Neighborhoods and health. New York: Oxford University Press; 2003.
2. Forsyth A, Lytle L, Van Riper D. Finding food: issues and challenges in using geographic information systems (GIS) to measure food access. *J Trans Land Use*. 2009 (in press).
3. Larson NI, Story MT, Nelson MC. Neighborhood environments: disparities in access to healthy foods in the U.S. *Am J Prev Med*. 2009; 36(1): 74-81.
4. Moore LV, Diez Roux AV, Nettleton JA, Jacobs DR Jr. Associations of the local food environment with diet quality—a comparison of assessments based on surveys and geographic information systems: the multi-ethnic study of atherosclerosis. *Am J Epidemiol*. 2008; 167(8): 917-924.
5. Brownson RC, Hoehner CM, Day K, Forsyth A, Sallis JF. Measuring the built environment for physical activity: state of the science. *Am J Prev Med*. 2009; 36(4 Suppl): S99–123 e112.

6. Diez Roux AV, Evenson KR, McGinn AP, et al. Availability of recreational resources and physical activity in adults. *Am J Public Health*. 2007; 97(3): 493-499.
7. Gordon-Larsen P, Nelson MC, Page P, Popkin BM. Inequality in the built environment underlies key health disparities in physical activity and obesity. *Pediatrics*. 2006; 117(2): 417-424.
8. Sallis JF, Hovell MF, Hofstetter CR, et al. Distance between homes and exercise facilities related to frequency of exercise among San Diego residents. *Public Health Rep*. 1990; 105: 179-185.
9. Kubzansky LD, Subramanian SV, Kawachi I, Fay ME, Soobader MJ, Berkman LF. Neighborhood contextual influences on depressive symptoms in the elderly. *Am J Epidemiol*. 2005; 162(3): 253-260.
10. Subramanian SV, Kubzansky L, Berkman L, Fay M, Kawachi I. Neighborhood effects on the self-rated health of elders: uncovering the relative importance of structural and service-related neighborhood environments. *J Gerontol B Psychol Sci Soc Sci*. 2006; 61 (3): S153-S160.
11. Bader MD, Ailshire JA, Morenoff JD, House JS. Measurement of the local food environment: a comparison of existing data sources. *Am J Epidemiol*. 2010; 171(5): 609-617.
12. Boone JE, Gordon-Larsen P, Stewart JD, Popkin BM. Validation of a GIS facilities database: quantification and implications of error. *Ann Epidemiol*. 2008; 18(5): 371-377.
13. Paquet C, Daniel M, Kestens Y, Leger K, Gauvin L. Field validation of listings of food stores and commercial physical activity establishments from secondary data. *Int J Behav Nutr Phys Act*. 2008; 5: 58.
14. Wang MC, Gonzalez AA, Ritchie LD, Winkleby MA. The neighborhood food environment: sources of historical data on retail food stores. *Int J Behav Nutr Phys Act*. 2006; 3: 15.
15. Melnick AL, Fleming DW. Modern geographic information systems—promise and pitfalls. *J Public Health Manag Pract*. 1999; 5(2): viii–x.
16. Forsyth A, Schmitz KH, Oakes M, Zimmerman J, Koepp J. Standards for environmental measurement using GIS: toward a protocol for protocols. *J Phys Act Health*. 2006; 3 (Suppl 1): S241-S257.
17. Handy SL, Clifton KJ. Evaluating neighborhood accessibility: possibilities and practices. *J Transport Stat*. 2001; 4: 67-78.
18. Porter DE, Kirtland KA, Neet MJ, Williams JE, Ainsworth BE. Considerations for using a geographic information system to assess environmental supports for physical activity. *Prev Chron Dis*. 2004; 1(4): A20.
19. Chao A, Tsay PK, Lin SH, Shau WY, Chao DY. The applications of capture-recapture models to epidemiological data. *Stat Med*. 2001; 20((20): 3123-3157.
20. Hook EB, Regal RR. Capture–recapture methods in epidemiology: methods and limitations. *Epidemiol Rev*. 1995; 17(2): 243-264.
21. Tilling K. Capture–recapture methods—useful or misleading? *Int J Epidemiol*. 2001; 30 (1): 12-14.
22. International Working Group for Disease Monitoring and Forecasting. Capture–recapture and multiple-record systems estimation II: applications in human diseases. *Am J Epidemiol*. 1995; 142(10): 1059-1068.
23. U.S. Census Bureau. *American Factfinder*. http://factfinder.census.gov. Accessed on: September 29, 2009.
24. ESRI. 2009 Methodology Statement: ESRI data—business locations and business summary. ESRI; 2009.
25. Dun and Bradstreet. *The DUNSRight (TM) process: the power behind quality information*. http://www.dnb.com/us/about/db_database/dnbinfoquality.html. Accessed on: September 29, 2009.
26. Vernez Moudon AV, Lee C, Cheadle AD, et al. Operational definitions of walkable neighborhood: theoretical and empirical insights. *J Phys Act Health*. 2006; 3(Suppl 1): S99-S117.

27. Forsyth A. Environmental and physical activity: GIS protocols. Vol Version 4.1 University of Minnesota and Cornell University; 2007. http://www.designforhealth.net/resources/gis_protocols.html. Accessed on: March 8, 2010.
28. Forsyth A, Oakes M, Schmitz KH, Hearst M. Does residential density increase walking and other physical activity? *Urban Stud*. 2007; 44(4): 679-697.
29. Baker EA, Schootman M, Barnidge E, Kelly C. The role of race and poverty in access to foods that enable individuals to adhere to dietary guidelines. *Prev Chron Dis*. 2006; 3(3): A76.
30. Zhao F, Gan A, Li S-C. Employment data comparison emphasizes importance of mulitiple data sources. *Fla Transp Model*. 1999; 12: 2-4.