DELAWARE CANCER BY CENSUS TRACT: METHODS

Geocoding Validation Process

When a cancer case is submitted to the Delaware Cancer Registry (DCR), geocoding software assigns the case to a census tract based on patient address at time of diagnosis. The accuracy of census tract assignment is entirely dependent on the accuracy and quality of patient address data. Several street address issues increase the likelihood of incorrect census tract assignment. For example, incorrectly spelled street names, multiple streets with the same name, incorrect or missing directional street information (e.g., North vs. South), and recently created streets that are not yet embedded within the geocoding software may result in inaccurate census tract assignment. Cancer cases with non-physical addresses (e.g., rural route and P.O. Boxes) are also difficult to accurately assign to census tracts.

Accurate census tract assignment is necessary for valid rate calculation at the census tract level. Therefore, prior to incidence rate calculation, the Delaware Division of Public Health (DPH) and the DCR conducted a multi-phase data validation process designed to verify that cancer cases had been accurately assigned to the census tract in which they were diagnosed.

The geocoding validation process included all cancer cases diagnosed in Delaware in 2002-2006. The first phase involved a case-level quality review of street address data. DCR staff began by correcting obvious street misspellings. Next, using Accurint®, a Lexis Nexis® service, DCR staff assigned a valid physical street address to P.O. Box addresses where possible. DCR staff also used Accurint to assign a valid physical street address to rural addresses where possible.

Next, DCR submitted a data file containing the remaining rural address cases to BCC Data Services. BCC Data Services utilized LACSLink software, which supports the conversion of rural addresses to street addresses. Using LACSLink, BCC Data Services successfully assigned valid physical street address data to an additional 66 cases with rural address data.

The second phase of the validation process focused on improving the accuracy of existing census tract data. Although the majority of DCR records had previously been assigned to a census tract when they were originally submitted to the DCR, the original census tract variable was associated with a fair degree of unreliability. To facilitate review of census tract data, the DCR contracted with Tele Atlas, a provider of digital map services. Tele Atlas performed a thorough check of the entire 2000-2006 cancer data file. Census tract data were corrected for N=3,950 cases that had previously been assigned to an incorrect census tract. Following receipt of the appended data file from Tele Atlas, the DCR conducted a final quality review of the file and updated their files accordingly.

Preliminary Analyses

Preliminary analyses were performed on one raw data file created for DPH by the DCR; the file included all cancer cases diagnosed in Delaware between January 1, 2002 and December 31, 2006 (N=24,854).

Per reporting guidelines mandated by the Surveillance, Epidemiology, and End Results (SEER) Program of the National Cancer Institute, cancer incidence rates *exclude* benign tumors, non-urinary bladder *in situ* tumors, and basal and squamous cell cancers. However, state cancer registries may still collect data on these tumors for tracking purposes. Therefore, the raw data file was analyzed to identify cases belonging to one of these categories. A total of 2,094 cases of benign tumors, non-urinary bladder *in situ* tumors, and basal and squamous cell cancers were eliminated from the file. Cases involving malignant tumors, as well as cases involving tumors with an unknown behavior code, were retained for further analyses.

The remaining 22,760 records were sorted by the variable “census tract certainty” (CTC). CTC codes are assigned using the following scale:

- **1 =** Census tract assignment was based on complete and valid physical street address
- **2 =** Census tract assignment was based on residence ZIP + 4
- **3 =** Census tract assignment was based on residence ZIP + 2
- **4 =** Census tract assignment was based on residence ZIP code only
-

Because they were assigned to a census tract based on complete street address data, cases with CTC=1 were considered to have the most accurate census tract data. Alternately, cases with a CTC > 1 were assigned to a census tract using the best available address data. If a physical street address was not available, the case was assigned to a census tract using a variation of the 5-digit ZIP code data. Cases with CTC=2 were considered to have the second-most accurate census tract data because census tract assignment was based on the extended ZIP + 4 code; this code provided a more precise geographic location than did the 5-digit ZIP code alone. Correspondingly, cases with CTC=3 or 4 were considered to have comparatively less accurate census tract data as census tract assignment was based on less specific variations of ZIP code data (ZIP+2 and ZIP code only, respectively).

A total of 707 cases had a CTC value > 1; these cases were pulled for individual review by DPH epidemiologists. Using publicly-available online map tools (e.g., Google Maps, MapQuest, and American Fact Finder (a service of the United States Census Bureau)), DPH epidemiologists re-assigned 55 of the 707 cases to a valid census tract; the CTC codes for these 55 cases were changed to 1 in the master data file. For another 53 cases, DPH epidemiologists confirmed the initial census tract assignment as accurate; the CTC codes for these cases were also changed to “1” in the master data file.

The remaining 599 records with CTC > 1 had incomplete or ambiguous address data, including P.O. Boxes, rural addresses, or unmappable street addresses. These 599 cases could not be accurately assigned to the census tract in which they were diagnosed; therefore, these cases were excluded from rate calculations at the census tract level. However, these 599 cases were retained for rate calculation at the state level. Table 1, below, shows the distribution of excluded cancer cases, by county of diagnosis.

Table 1: Excluded Cancer Cases, by County of Diagnosis

County	**Number of Cases Excluded**
New Castle	71
Kent	57
Sussex	471
Total	***599***

Note that Sussex County was disproportionately represented in the 599 cases excluded from census tract analyses. As a result, 2002-2006 incidence rates for Sussex County census tracts are suppressed to a greater extent than are incidence rates for census tracts in New Castle and Kent Counties.

In the future, fewer cases will be excluded from rate calculations due to unmappable street addresses. The DCR has developed new protocol to ensure that when cases with P.O. Box and rural addresses are first reported to the registry, they are immediately flagged for additional follow-up and supplementation with physical address data.

Calculating Five-Year Population Estimates, by Census Tract

Delaware is subdivided into 197 census tracts. Note that census tracts do not follow a consecutive numbering scheme. New Castle County includes tracts 1.00 through 169.02. Kent County is comprised of tracts 401.00 through 431.00, and Sussex County includes tracts 501.01 through 519.00.

Census tract populations were calculated using estimates from the Delaware Population Consortium (DPC) and the 2000 Census. DPC census tract population estimates were available for all years 2002 through 2006. DPH staff used 2000 Census data to calculate the proportion that each 5-year age group contributed to the overall census tract population. These proportions were applied to DPC-based census tract estimates to yield annual population estimates by census tract, broken down by 5-year age groups. DPH staff repeated this process for male and female populations to obtain gender-specific 2002-2006 population estimates by 5-year age groups for each census tract.

Denominators for years 2002 through 2006 were summed to obtain the 2002-2006 population for each census tract. Five-year (2002-2006) population estimates ranged in size from 3,132 for Census Tract 10 to 65,136 for Census Tract 148.06. Both of these census tracts are located in New Castle County.

Calculating Age-Adjusted Incidence Rates, by Census Tract

Census tract-level incidence rates were calculated from a modified dataset including N=22,161 cases diagnosed between 2002 and 2006[1]. Within the cancer data file, cross-tabulations (age group x census tract) were performed to determine the number of cancer cases diagnosed by census tract and the age groups in which they were diagnosed. These frequencies were used to calculate crude and age-adjusted incidence rates at the census tract level. Crude incidence rates represent the total number of new cancer diagnoses divided by the total population at risk, without consideration of any demographic characteristics of the population. Age-adjusted incidence rates take into account the age distribution of the population at risk; age-adjusted incidence rates are useful for comparing rates between two populations that differ in age composition.

To calculate crude incidence rates, the number of cancer cases diagnosed in a particular age group in a particular census tract was divided by the population size for that specific cohort; these values were then multiplied by 100,000 (see Equation 1). To determine the 2002-2002 crude incidence rate for an entire census tract, the number of cancer cases diagnosed in a census tract over the 5-year period was divided by the total population of the census tract for the same 5-year period, and this value was multiplied by 100,000.

Equation 1: 2002-2002 Crude All Site Incidence Rate, 40-44 year olds, Census Tract 425.00

$$\frac{(\text{No. cancer cases } (2002-2006) \text{ among } 40-44 \text{ year olds in CT}425.00)}{(2002-2006 \text{ population, } 40-44 \text{ year olds in CT}425.00)} = \frac{(2)}{(1195)} \times 100{,}000 = \quad 167.3 \text{ per } 100{,}000$$

To calculate age-adjusted incidence rates, crude incidence rates for each age group were multiplied by the appropriate 2000 U.S. Standard Million Population weight[2]. Table 2 displays the U.S. Standard Million population weights, by age group. Age-adjusted incidence rates for each of the 18 age groups were summed to yield the age-adjusted incidence rate for an entire census tract.

Table 2: U.S. Standard Million Population Weights, by Age Group

Age Group	**U.S. Standard Million Population Weight**
0-4 yrs	0.0691
5-9 yrs	0.0725
10-14 yrs	0.0730
15-19 yrs	0.0722
20-24 yrs	0.0665
25-29 yrs	0.0645
30-34 yrs	0.0710
35-39 yrs	0.0808
40-44 yrs	0.0819
45-49 yrs	0.0721
50-54 yrs	0.0627
55-59 yrs	0.0485
60-64 yrs	0.0388
65-69 yrs	0.0343
70-74 yrs	0.0318
75-79 yrs	0.0270
80-84 yrs	0.0178
85+ yrs	0.0155

Source: Centers for Disease Control and Prevention,
National Center for Health Statistics

[1] The modified sample size reflected the N=599 cases that were eliminated from census tract-level analyses because they could not accurately be assigned to the census tract in which they were diagnosed.

[2] Published by the Centers of Disease Control and Prevention and the National Center for Health Statistics.

Calculating the Age-Adjusted Incidence Rate for the State of Delaware

The average annual age-adjusted cancer incidence rate for the state of Delaware was calculated from the full dataset including N=22,161 cases diagnosed between 2002 and 2006. Cross-tabulations (age group x census tract) were performed to determine the number of cancer cases diagnosed in the state and the age groups in which they were diagnosed. Using the process described above, frequencies were used to calculate crude and age-adjusted incidence rates at the state level.

Calculating 95% Confidence Intervals

Confidence intervals represent the range of values in which the cancer rate could reasonably fall. Our best estimate of the cancer rate in a particular census tract is the incidence rate, itself. However, the rate could reasonably lie anywhere between the lower confidence limit (LCL) and the upper confidence limit (UCL). Because of this, a confidence interval is sometimes called the "margin of error." Confidence intervals were calculated for all census tract-level incidence rates, as well as for the state incidence rate.

When incidence rates were based on more than 100 cases, 95% confidence intervals were calculated using the following formulas:

$$\text{Lower Confidence Limit} = \quad \text{AA Rate} - 1.96\left(\frac{(\text{AA Rate})}{\sqrt{\#\ \text{Cases}}}\right)$$

$$\text{Upper Confidence Limit} = \quad \text{AA Rate} + 1.96\left(\frac{(\text{AA Rate})}{\sqrt{\#\ \text{Cases}}}\right) ,$$

where AA Rate = the age-adjusted incidence rate for a particular census tract.

When incidence rates were based on fewer than 100 cases, 95% confidence intervals were calculated using the following formulas:

$$\text{Lower Confidence Limit} = \quad \text{AA Rate x L}$$

$$\text{Upper Confidence Limit} = \quad \text{AA Rate x U} ,$$

where AA Rate = the age-adjusted incidence rate for a particular census tract, and L and U = values published by the National Center for Health Statistics for the specific purpose of calculating 95% confidence intervals for rates computed from fewer than 100 cases[3].

Comparing Census Tract Rates to the State Rate

The level of uncertainty associated with an incidence rate is reflected in the width of its confidence interval. Very wide confidence intervals mean that the incidence rate is estimated with a large degree of uncertainty.

The width of a confidence interval is influenced by two factors: (a) the number of cancer cases in the population under consideration and (b) the size of the population under consideration. When a cancer rate is calculated for a small population in which only a handful of cases were diagnosed, we would expect the confidence interval for the rate will be very wide. On the other hand, when a cancer rate is calculated for a very large population in which many cases were diagnosed, we would expect the confidence interval for the rate will be very narrow.

The width of a confidence interval is important because it is used to determine if the amount by which two incidence rates differ is statistically significant. If the confidence interval for the incidence rate in one area overlaps with the confidence interval for an incidence rate in another area, the rates are not significantly different from one another. Researchers interpret a non-significant difference as "no meaningful difference" between rates. Even though the two rates may look very different, if the cancer rate for one

[3] Martin JA, Hamilton BE, Ventura SJ, Menacker F, Park MM, Sutton PD. Births: Final data for 2001. National vital statistics reports; vol 51 no. 2. Hyattsville, Maryland: National Center for Health Statistics. 2002.

area is NOT significantly different from the cancer rate for another area, researchers cannot say that one rate is truly different from the other rate.

On the other hand, if the confidence interval for the incidence rate in one area does NOT overlap with the confidence interval for an incidence rate in another area, the two rates are significantly different. When the rate for one area is significantly different from the rate for another area, the difference between the rates is larger than would be expected by chance alone.

DPH compared the all-site incidence rate for each census tract to the all-site incidence rate for the state of Delaware. This allowed DPH to identify any census tracts with incidence rates that are higher or lower than the incidence rate for Delaware as a whole. If the confidence interval for a census tract incidence rate overlapped with the confidence interval for the state incidence rate, the census tract rate was not significantly different from the state rate. If the confidence interval for a census tract rate did not overlap with the confidence interval for the state rate, the census tract rate was significantly different from the state rate. Census tracts with significantly higher or lower cancer rates compared to the state are denoted in the rate table and all color-coded maps.

Supplemental Information

Incidence rates for nine census tracts were based on fewer than 25 cases. When incidence rates are computed for an entire geographic area based on a very small number of cases, rates are estimated with a larger degree of uncertainty. This uncertainty is represented by a very wide confidence interval. When confidence intervals are wide, they are more likely to overlap with the confidence intervals of incidence rates from other areas; this means that it is more difficult to establish a significant difference between incidence rates. For this reason, rates based on fewer than 25 cases should especially be interpreted with caution. To assist interpretation, incidence rates calculated from fewer than 25 cases are denoted in both the rate table and color-coded maps.