NEW MEXICO ENVIRONMENTAL PUBLIC HEALTH TRACKING PROGRAM

Short-Latency Cancer Project

Results of Case Address Review and Geocoding

Submitted by

William F. Athas Ph.D.
Research Assistant Professor
Department of Family and Community Medicine
University of New Mexico School of Medicine
Albuquerque, NM 87131

December, 2005

Introduction

Environmental tracking of cancer is problematic due to long latency periods and the highly mobile nature of the US population. One exception are cancers occurring in children, many of which present within the first several years of life, thus providing a somewhat more "geographically stable" outcome compared to adult cancers, where latencies may range from 20-50 years. The most common childhood cancers are hematopoetic tumors, principally leukemias, and to a lesser extent lymphomas. Brain-related tumors are second most common in children, many of which present as malignant tumors of glial origin. Another potential "environmentally-visible" cancer is thyroid cancer, a common tumor among women of reproductive age, and one which has undergone a substantial unexplained increase in incidence in both sexes over the past two decades. The development of clinical thyroid cancer may be influenced by late-stage factors - it is a slow-growing, generally indolent cancer which appears to be harbored in relatively high population prevalence as a sub-clinical form. Late-stage agents promoting transition of sub-clinical to clinical thyroid cancer may be environmental in origin, and thus amenable to study through environmental health surveillance. This report provide results from data development activities in support of establishing an environmental public health surveillance activity for such short-latency cancers in New Mexico. The scope of the report is limited to data quality improvement and geocoding methods.

Methods

Data Source

Data on short-latency cancers were obtained from a confidential master case file provided to the NM-EPHT program by the New Mexico Tumor Registry (NMTR). The master file contained all individual case records contained in the NMTR Case Information Management System (CIMS) as of January 18, 2005 for the 15-year diagnosis time period 1988-2002. Included in the file was information on select tumor characteristics, non-confidential patient demographics, and address at diagnosis. (The NMTR file lacked addresses on American Indian cases, and they were thus excluded from analysis.) Short-latency cancers were selected based on topology and histology codes defined by the National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) program and grouped as follows:

- Leukemia (0-19 Years)
- Lymphoma (0-19 Years)
- Non-hematopoetic cancers (0-19 Years)
- Thyroid cancer (All ages)
- Brain-related tumors[1] (All ages)

[1] Brain-related tumors refer to any neoplasm of the brain, central nervous system, pituitary gland, craniopharyngeal duct, or pineal gland, regardless of tumor behavior, and includes malignant, benign, and tumors of uncertain behavior. Although statewide population-based collection of malignant brain tumors has been on-going since 1969, cases of benign and uncertain behavior have only been collected on a population basis since 1996.

Assignment of Address Type Indicator

Case records for the short-latency cancers were entered into an analytic file, sorted on the address at diagnosis field, and assigned an address type code based on the following coding scheme:

Table 1: Data Dictionary for Address Type Field

ADDRESS TYPE	CODE
Complete street address	01
Street name only (+ City/Zipcode)	02
Intersection	03
Rural route \ Route\ Star route	04
Highway Contract Route (HCR)	05
PO Box \ Box	06
General Delivery	07
Physical location only	08
Military facility	09
Correctional facility	10
Other facility (Nursing home, ranch, etc)	11
City and Zipcode only	12
City only	13
NM only	99

Some records contained multiple address types in the address field, such as a PO box and a numbered street address. In such instances, address type was assigned based on the entry containing the greatest amount of geospatial detail.

Case Address Review

Case records assigned address codes 02 - 99 were selected and submitted for manual address tracing using a variety of different resources, including NMTR case records, hardcopy telephone and reverse directories, and free on-line mapping and tracing resources. Where possible, case addresses were upgraded using information obtained from the tracing resources, when the information was deemed accurate and timely, that is, current to within a year of the date of diagnosis. Address tracing employed personal identifying information, including name, social security number, age, and date at diagnosis, and required on-site access to confidential NMTR case records, which was granted under agreement between NM-EPHT and NMTR as part of the on-going surveillance activities of NMTR, and approved as such by the UNM Human Research Review Committee (HRRC). In order to accommodate the upgraded case address information, the original address field was copied into a new field where all upgrades, edits, changes, and modifications were made, thus leaving the original field intact for review and reference purposes. A second address type field was added to capture the address characteristics of the new “modified ” address field.

Geocoding: Latitude-Longitude Coordinate Assignment

Figure 1 shows the geocoding procedures used to assign latitude-longitude coordinates to case records. Various different methods were used depending on the address type, as described below. Following assignment of all lat-long coordinates, the spatial join function in ArcGIS 9.1 was used to place the coordinates within census tracts (Census 2000) in order to enable data linkage and analysis at the tract level.

Address matching for complete street addresses

Cases records with a complete street address were submitted to address clean-up software (Semaphore ZP4, June 2005) in order to standardize the addresses to the US Postal Service (USPS) format. Address standardization increases both the number of address matches as well as address match scores. Following address clean-up, case records were address matched using ArcGIS 9.1 software with the ESRI StreetMap 2005 street reference file. StreetMap 2005 is referenced to the North American Datum of 1983 (NAD83) with it's address elements prestandardized based on ESRI address standardization rules, which generally conform to the USPS format. Geocoding parameters included a spelling sensitivity of 80, 3% inset, and 50 foot offset. The offset was needed in order to draw the points off the street centerline, thereby providing a better approximation of reality, as well as improving the accuracy of spatially joining geocoded coordinate points with area (polygon) features, such as census tracts. All records with a match score of less than 100 (match range 0-100, where 100 indicates complete agreement on all standardized address elements) were submitted for interactive re-matching through several iterations interspersed with manual address review. Case addresses which could not be edited or otherwise modified to provide a match score of 100 were submitted for manual coordinate assignment.

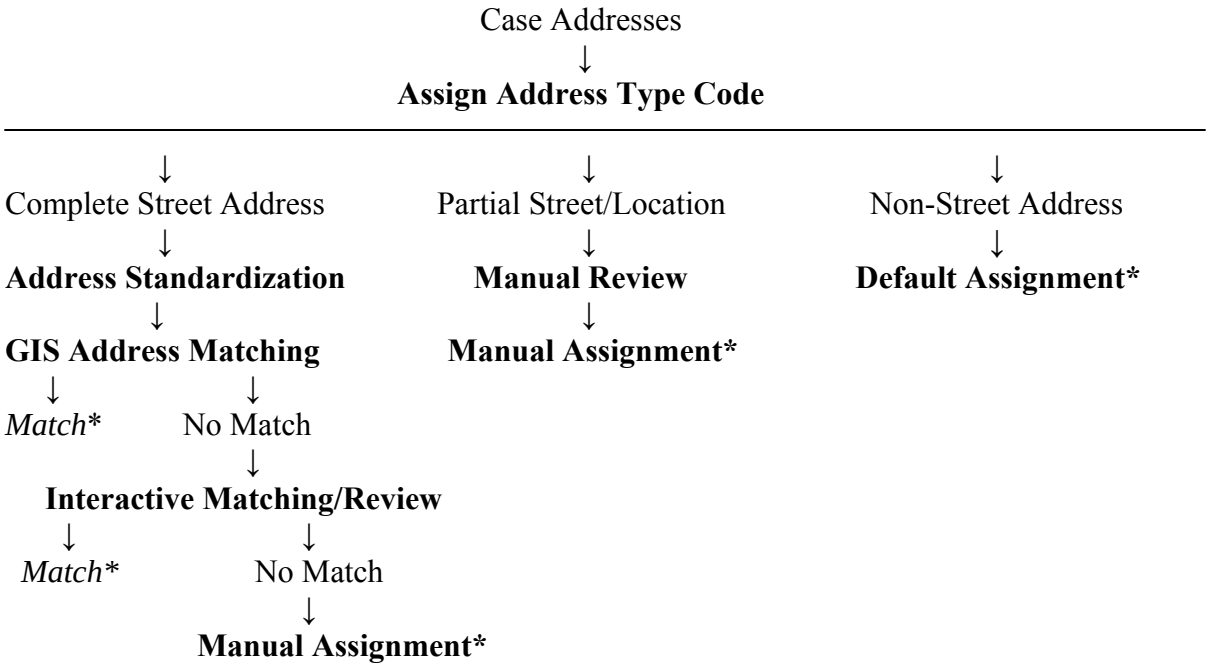
Manual coordinate assignment for unmatched street addresses and partial addresses

Manual coordinate assignment for unmatched complete and partial street addresses was performed using ESRI StreetMap 2005 visualized in ArcMap in conjunction with on-line mapping services, including Mapquest, Google, and Yahoo maps. Those complete street addresses which failed to achieve a match score of 100 typically had an out-of-range address or a street name not identified in the street reference file. For such addresses, on-line mapping services were first queried, followed by basic internet searches to identify the most likely location of the address. The approximated address location was visualized on StreetMap 2005 and a lat-long coordinate point assigned manually.

Default geocoding for non-street addresses

Non-street addresses were geocoded using one of three principal methods: (1) census tract centroid; (2) post office location; or (3) Geographic Names Information System (GNIS)[2] feature centroid. Census tract centroids were used primarily in the Santa Fe area for rural route addresses, which are still commonly used and which to some extent can be geographically localized, at least roughly to a census tract level, using on-line search

Figure 1. Geocoding Scheme Used to Assign Latitude-Longitude Coordinates



* Latitude-longitude coordinate assignment

engines and resources. Post office location was used primarily for PO box addresses in the Albuquerque area, where multiple post offices spread throughout the city service the various different address ranges and zip codes associated with PO boxes. Here the assumption is that it is more likely than not that one lives closest to the post office servicing their PO box. For the majority of non-street addresses, many of which localize to the small and mid-size towns that populate much of New Mexico, GNIS coordinates were assigned based on feature name, typically a populated place (e.g., city) or locale (e.g., ranch). Although geocoding to Zip Code centroid is a possible alternative method, GNIS-based feature centroids are considered superior given the rural nature of New Mexico, where multiple towns may exist within a single zipcode, and where the centroid of a zip code may be at some distance from the nearest population center. Furthermore, zip code boundary files are neither standardized nor maintained by the USPS, despite the addition of new zip codes with population growth.

[2] *Geographic Names Information System (GNIS)* - the database developed by the USGS in cooperation with the U.S. Board on Geographic Names which contains the federally-recognized geographic names for all known places, features, and areas in the United States that are identified by a proper name. Each feature is located by state, county, and geographic coordinates and is referenced to the appropriate 1:24,000-scale or 1:63,360-scale USGS topographic map on which it is shown.

Non-geocodable case records

A small number of cases were entered in the NMTR file based solely on state of residence, that is, no address information was available other than the case was a resident, or presumed resident, of New Mexico at the time of diagnosis.

Geocoding: Census Tract Assignment

Case records were assigned a residence census tract at time of diagnosis using the spatial join function in ArcGIS 9.1. In brief, lat-long coordinates were spatially joined with census tract polygons (Census 2000) using point-in-polygon geometry contained within ArcGIS. Since addresses whose coordinates are placed on street centerlines that coincide with tract boundaries (as well as boundaries of other administrative areas) can be incorrectly assigned, all address matching was performed using a 50-foot offset, as previously noted above. Furthermore, experience has shown that census tract boundary files may contain spatial inaccuracies, that throw a boundary off the street centerline, in which case, an offset again promotes more accurate tract assignment.

Geocoding: Quality Assurance and Control

Positional Accuracy

Positional accuracy refers to the difference between the true and assigned coordinates of a feature. No attempt was made here to assess the degree of positional accuracy of the assigned lat-long coordinates. Such assessments require “gold-standard” coordinates, as may be obtained through Global Positioning System (GPS) measurements or other spatially accurate measures. Furthermore, the absolute relevance of positional accuracy

will vary depending on the intended use of the geocoded data. For example, where the intent is to conduct individual level analyses involving proximity measures, positional inaccuracy may severely limit the interpretation and value of a study. On the other hand, where the intent simply is to place a coordinate within a polygon, such as a tract, the analyses may be much less sensitive to positional inaccuracy. In such cases, the importance and potential relevance of positional inaccuracy is directly related to the areal size of the polygon into which coordinates are placed. Using vetted national street reference files, such as StreetMap 2005, minimizes the possibility that major positional inaccuracies are incurred during address matching.

Assignment Accuracy

Assignment accuracy refers to the difference between the true and assigned area into which an address is placed, such as a census tract, block group, administrative area, or exposure polygon. For administrative units, assignment accuracy can be assessed by comparing the assigned area with that obtained through independent geocoding sources, such as that provided by the US Census. Prior results geocoding 1988-2002 New Mexico bladder cancer cases, a dataset which underwent 100% review of census tract assignments, showed that only a small number of address matched points were assigned into an incorrect census tract. Of those, virtually all occurred in close proximity to tract boundary lines, and involved either errors in the StreetMap 2005 reference file or spatial inaccuracy in tract boundary lines. Consequently, quality assurance measures for tract assignment of short latency cancers focused on those coordinate points falling in proximity to tract boundary lines. The buffer function within ArcGIS was used to select all coordinate points falling within 100 ft of a tract boundary, and this subset of points then submitted to two independent geocoding sources: the US Census (http://www.census.gov) and the Federal Financial Institutions Information Council (FFIEC) (http://www.ffiec.gov/geocode/). Discrepancies between assigned tract and that returned by the two geocoding services were reconciled and entry made to document the underlying cause of incorrect tract assignments.

Results

Address Type in 1988-2002 NM Short-Latency Cancers

Table 2 shows the distribution of the 1988-2002 short-latency cancers according to address type and site. Complete street addresses were by far the dominant address type for each cancer site, ranging from 78% for non-hematopoetic childhood cancers to 82.9% for brain-related tumors. The second most common address type was a PO box, which also was similarly distributed across cancer sites, ranging from 13.6% for brain-related tumors to 17.9% for non-hematopoetic childhood cancers. The third most frequent address type was rural route, which averaged 2% of addresses across the different cancer sites. The remaining address types accounted for less than 5% of the cases, regardless of cancer site. Of note, was the virtual absence of records containing simply a general location descriptor within the address field.

Table 2. Frequency Distribution of Address Type: 1988-2002 NM Short-Latency Cancers

ADDRESS TYPE	Cancer Cases (0-19 Years)						Cancer Cases (All Ages)			
	Leukemia		Lymphoma		Non-Hematopoetic		Thyroid Cancer		Brain-Related Tumors	
	Cases	%	Cases	%	Cases	%	Cases	%	Cases	%
Complete street address	248	79.2%	86	78.9%	514	78.0%	1,431	81.8%	1,774	82.9%
Street name only	2	0.6%	0	0.0%	3	0.5%	2	0.1%	2	0.1%
Intersection	0	0.0%	0	0.0%	0	0.0%	1	0.1%	0	0.0%
Rural route/route/star route	6	1.9%	2	1.8%	14	2.1%	35	2.0%	48	2.2%
Highway contract route	1	0.3%	1	0.9%	5	0.8%	20	1.1%	16	0.8%
PO Box	48	15.3%	19	17.4%	118	17.9%	244	14.0%	290	13.6%
General Delivery	0	0.0%	0	0.0%	1	0.2%	3	0.2%	1	0.1%
Military facility	0	0.0%	0	0.0%	0	0.0%	2	0.1%	0	0.0%
Corrections facility	0	0.0%	0	0.0%	0	0.0%	5	0.3%	0	0.0%
Other facility	0	0.0%	0	0.0%	0	0.0%	1	0.1%	0	0.0%
City and zip code only	1	0.3%	0	0.0%	1	0.2%	2	0.1%	4	0.2%
City only	0	0.0%	0	0.0%	0	0.0%	1	0.1%	3	0.1%
NM only	7	2.2%	1	0.9%	3	0.5%	2	0.1%	3	0.1%
TOTAL	313	100.0%	109	100.0%	659	100.0%	1,749	100.0%	2,141	100.0%

Geocoding

Assignment of latitude-longitude coordinates

Tables 3 and 4 show geocoding results according to cancer site. Rates of address matching (match score = 100) of complete street addresses were similar across sites, ranging from 92.4% to 95.2%. Virtually all of the complete street addresses which failed automated and interactive matching were geocoded by manual placement of lat-long coordinates. Assignment of lat-long coordinates based on GNIS centroids was the most common method used for non-street addresses.

Tables 5-7 show the geographic variability in the percent of case records geocoded by address matching or manual assignment - methods based on the actual location of the case residence at diagnosis, and thus the most accurate geospatial information available. Considerable geographic variation in the percent of geocodes based on actual street address was observed. Among towns with 10 or more cases, the percentage ranged from less than 20% up to 100%, a pattern seen at each cancer site. The Albuquerque area, which typically accounts for roughly a third of all state cancer cases, consistently had geocodes based on street location exceeding 97%. In sharp contrast, among the more rural areas of the state, as represented collectively by the towns with less than 10 total cases (i.e., Rest of State), the percent of geocoding based on street information was less than 50%; regardless of cancer site.

Table 3. Distribution of Geocode by Address Type and Site: Childhood Cancers

Address Type	Type of Geocode											
	Address Match (Score = 100)		Manual Assignment		GNIS Centroid		Tract Centroid		Post Office Location		Not Geocodable	
	Cases	Row %	Cases	Row %	Cases	Row %	Cases	Row %	Cases	Row %	Cases	Row %
LEUKEMIA (0-19 Yrs)												
Complete street address	236	95.2%	9	3.6%	2	0.8%	0	0.0%	1	0.4%	0	0.0%
Street name only	0	0.0%	0	0.0%	2	100.0%	0	0.0%	0	0.0%	0	0.0%
Rural route/route/star route	0	0.0%	0	0.0%	4	66.7%	1	16.7%	1	16.7%	0	0.0%
Highway contract route	0	0.0%	0	0.0%	1	100.0%	0	0.0%	0	0.0%	0	0.0%
PO Box	0	0.0%	0	0.0%	47	97.9%	0	0.0%	1	2.1%	0	0.0%
City and zip code only	0	0.0%	0	0.0%	1	100.0%	0	0.0%	0	0.0%	0	0.0%
NM only	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	7	100.0%
LYMPHOMA (0-19 YRS)												
Complete street address	80	93.0%	6	7.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%
Rural route/route/star route	0	0.0%	0	0.0%	2	100.0%	0	0.0%	0	0.0%	0	0.0%
Highway contract route	0	0.0%	0	0.0%	1	100.0%	0	0.0%	0	0.0%	0	0.0%
PO Box	0	0.0%	0	0.0%	18	94.7%	0	0.0%	1	5.3%	0	0.0%
NM only	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	1	100.0%
NON-HEMATOPOETIC CANCERS (0-19 Yrs)												
Complete street address	475	92.4%	36	7.0%	3	0.6%	0	0.0%	0	0.0%	0	0.0%
Street name only	0	0.0%	1	33.3%	2	66.7%	0	0.0%	0	0.0%	0	0.0%
Rural route/route/star route	0	0.0%	0	0.0%	13	92.9%	1	7.1%	0	0.0%	0	0.0%
Highway contract route	0	0.0%	0	0.0%	5	100.0%	0	0.0%	0	0.0%	0	0.0%
PO Box	0	0.0%	0	0.0%	113	95.8%	0	0.0%	5	4.2%	0	0.0%
General Delivery	0	0.0%	0	0.0%	1	100.0%	0	0.0%	0	0.0%	0	0.0%
City and zip code only	0	0.0%	0	0.0%	1	100.0%	0	0.0%	0	0.0%	0	0.0%
NM only	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	3	100.0%

Table 4. Distribution of Geocode by Address Type: Brain-Related Tumors & Thyroid Cancer

Address Type	Type of Geocode											
	Address Match (Score = 100)		Manual Assignment		GNIS Centroid		Tract Centroid		Post Office Location		Not Geocodable	
	Cases	Row %	Cases	Row %	Cases	Row %	Cases	Row %	Cases	Row %	Cases	Row %
Brain-Related Tumors (All Ages)												
Complete street address	475	92.4%	36	7.0%	3	0.6%	0	0.0%	0	0.0%	0	0.0%
Street name only	0	0.0%	1	33.3%	2	66.7%	0	0.0%	0	0.0%	0	0.0%
Rural route/route/star route	0	0.0%	0	0.0%	13	92.9%	1	7.1%	0	0.0%	0	0.0%
Highway contract route	0	0.0%	0	0.0%	5	100.0%	0	0.0%	0	0.0%	0	0.0%
PO Box	0	0.0%	0	0.0%	113	95.8%	0	0.0%	5	4.2%	0	0.0%
General Delivery	0	0.0%	0	0.0%	1	100.0%	0	0.0%	0	0.0%	0	0.0%
City and zip code only	0	0.0%	0	0.0%	1	100.0%	0	0.0%	0	0.0%	0	0.0%
NM only	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	3	100.0%
Thyroid Cancer (All Ages)												
Complete street address	1,337	93.4%	92	6.4%	0	0.0%	2	0.1%	0	0.0%	0	0.0%
Street name only	0	0.0%	2	100.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%
Intersection	0	0.0%	1	100.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%
Rural route/route/star route	0	0.0%	1	2.9%	3	8.6%	25	71.4%	6	17.1%	0	0.0%
Highway contract route	0	0.0%	0	0.0%	0	0.0%	20	100.0%	0	0.0%	0	0.0%
PO Box	0	0.0%	0	0.0%	24	9.8%	220	90.2%	0	0.0%	0	0.0%
General Delivery	0	0.0%	0	0.0%	0	0.0%	3	100.0%	0	0.0%	0	0.0%
Military facility	0	0.0%	0	0.0%	0	0.0%	0	0.0%	2	100.0%	0	0.0%
Corrections facility	0	0.0%	5	100.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%
Other facility	0	0.0%	1	100.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%
City and zip code only	0	0.0%	0	0.0%	1	50.0%	1	50.0%	0	0.0%	0	0.0%
City only	0	0.0%	0	0.0%	0	0.0%	1	100.0%	0	0.0%	0	0.0%
NM only	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	2	100.0%

Table 5. Distribution of Address Matched and Manually Assigned Geocodes According to Geographic Area (10+ Total Cases): Childhood Cancers (0- 19 Years)

GEOGRAPHIC AREA	Total Cases	Address Match & Manual Assignment	
		No.	%
NEW MEXICO	1081	843	78.0
HOBBS	30	30	100.0
RIO RANCHO	30	30	100.0
CARLSBAD	15	15	100.0
ALBUQUERQUE	358	349	97.5
CLOVIS	21	20	95.2
ALAMOGORDO	18	17	94.4
ROSWELL	40	36	90.0
DEMING	10	9	90.0
LAS CRUCES	68	61	89.7
FARMINGTON	24	21	87.5
SANTA FE	50	42	84.0
LAS VEGAS	17	14	82.4
GALLUP	14	10	71.4
LOS LUNAS	16	11	68.8
BELEN	10	6	60.0
BERNALILLO	10	6	60.0
ANTHONY	20	11	55.0
ESPANOLA	10	5	50.0
SILVER CITY	10	5	50.0
LOVINGTON	10	4	40.0
REST OF STATE	289	141	48.8

Table 6. Distribution of Address Matched and Manually Assigned Geocodes According to Geographic Area (10+ Total Cases): Thyroid Cancer (All Ages)

GEOGRAPHIC AREA	Total Cases	Address Match & Manual Assignment	
		No.	%
NEW MEXICO	1749	1439	82.3
ROSWELL	38	38	100.0
HOBBS	31	31	100.0
ALAMOGORDO	22	22	100.0
PORTALES	19	19	100.0
WHITE ROCK	19	19	100.0
LAS VEGAS	14	14	100.0
RIO RANCHO	65	64	98.5
ALBUQUERQUE	691	680	98.4
CLOVIS	24	23	95.8
CARLSBAD	47	45	95.7
LAS CRUCES	70	66	94.3
FARMINGTON	30	27	90.0
ARTESIA	10	9	90.0
RATON	17	15	88.2
LOS ALAMOS	30	26	86.7
BELEN	12	10	83.3
SANTA FE	109	89	81.7
SOCORRO	14	11	78.6
LOS LUNAS	22	17	77.3
LOVINGTON	13	10	76.9
SILVER CITY	12	9	75.0
GRANTS	12	8	66.7
CORRALES	13	8	61.5
ANTHONY	13	7	53.8
ESPANOLA	12	6	50.0
RUIDOSO	15	7	46.7
EDGEWOOD	10	4	40.0
BERNALILLO	13	2	15.4
REST OF STATE	352	153	43.5

Table 7. Distribution of Address Matched and Manually Assigned Geocodes According to Geographic Area (10+ Total Cases): Brain-Related Tumors (All Ages)

GEOGRAPHIC AREA	Total Cases	Address Match & Manual Assignment	
		No.	%
NEW MEXICO	2141	1772	82.8
TRUTH OR CONSEQUENCES	16	16	100.0
ALBUQUERQUE	778	774	99.5
RIO RANCHO	53	52	98.1
HOBBS	37	36	97.3
LAS CRUCES	137	132	96.4
FARMINGTON	61	58	95.1
AZTEC	17	16	94.1
ARTESIA	25	23	92.0
ALAMOGORDO	37	34	91.9
GRANTS	12	11	91.7
CARLSBAD	46	42	91.3
CLOVIS	45	41	91.1
ROSWELL	80	72	90.0
PORTALES	16	14	87.5
RATON	13	11	84.6
LOS LUNAS	23	19	82.6
TULAROSA	11	9	81.8
SANTA FE	125	102	81.6
TUCUMCARI	16	13	81.3
BELEN	20	16	80.0
LOS ALAMOS	10	8	80.0
DEMING	30	23	76.7
SILVER CITY	26	19	73.1
GALLUP	14	10	71.4
BLOOMFIELD	12	8	66.7
LAS VEGAS	11	7	63.6
LOVINGTON	21	13	61.9
ANTHONY	21	12	57.1
EDGEWOOD	18	9	50.0
ESPANOLA	12	4	33.3
CORRALES	10	3	30.0
MORIARTY	10	3	30.0
TAOS	12	2	16.7
REST OF STATE	366	160	44.0

Assignment of census tracts

A GIS spatial join function was used to place latitude-longitude coordinates in Census 2000 tracts using tract boundary files downloaded from ESRI Geography Network. Following the spatial join, all address-matched points located within 100 feet of a tract boundary were selected using a GIS buffer function and submitted for independent tract geocoding. The results of this quality assurance step showed near complete agreement with tract assignments made under the geocoding protocol described here and the assignments returned from two on-line geocoding services. The small number of discrepancies which did arise were found to be the result of either errors in the street reference file or geospatial inaccuracies in the tract boundary file – attributes which can be corrected. Thus, with further fine tuning of the data resources used in the geocoding step, there is reason to believe that assignment of coordinate points into census tracts, or other geospatially defined administrative units, can proceed essentially error-free. This of course, does not relinquish the burden of rendering the most accurate geospatial information available for a given health record in order to place coordinate points at a location that best reflects reality.

Conclusions

New Mexico has had statewide, population-based cancer registration on-going continuously since 1969. The New Mexico Tumor Registry (NMTR), the state agency responsible for operating a cancer registry in New Mexico, has been a member state of the NCI SEER program since 1973, and as such, has been required to conform its data collection and quality assurance procedures according to the strict standards imposed by the SEER program. Thus, cancer incidence data in New Mexico are generally considered of high quality, particularly in regards to information collected on tumor characteristics and patient demographics. However, being a largely rural state, and one in which use of non-street mailing addresses in not uncommon, the geospatial information contained in state cancer registry records varies considerably. This variability is in large part due to the absence of formal address tracing at the time of initial case registration and data abstraction.

The results of the data enhancements and geocoding performed here on 1988-2002 short latency cancers shows that once outside the Albuquerque metro area, the percentage of records geocoded based on actual street parameters drops off fairly quickly, particularly in rural areas away from the state's main population centers. The geospatial accuracy of the cancer records shows a strong nodal pattern, where the more accurate geospatial case information is clustered in towns and cities with well-defined street architectures and modest population sizes. This nodal pattern will bear directly on the type and nature of environmental health surveillance activities mounted in New Mexico towards malignant diseases. For example, use of routine cluster analysis tools at the state level may not be feasible given the prevalent use of centroids for geocoding in rural areas. Such practice tends to cluster cases artificially. Cluster analysis may only be feasible in geographically contiguous areas of high quality geospatial information, such as the Albuquerque metro area. Similarly, the geographic unit of analysis may need to be varied and expanded

beyond census tract in order to avoid substantial misclassification when linking case numerator data with appropriate population denominator data. The same holds for area-level linkage with environmental determinants, particularly when such determinants display a complicated geographic pattern of hazard or exposure necessitating large-scale analysis.

A principle goal of all surveillance programs is continual data quality improvement. Given the heterogeneous nature of the geospatial information available for geocoding incident cancer cases in New Mexico, efforts should be directed and assistance provided where possible to prospectively improve the quality of the case address field at the time of initial case registration and data abstraction. Experience with the statewide birth registration system has shown that such improvements are both feasible and attainable to the point of near complete population of address fields with street-based or locational parameters.