Characterizing Anticipated Mobility
# of the NCS Cohort

by

Jeff Lehman, Jyothi Nagaraja, and Warren Strauss

1. Introduction

The U.S. Census Bureau estimates that roughly 46 percent of all persons aged five years or older have moved at least once between 1995 and 2000. Moreover, young adults between 25 and 39 years of age represented the largest segment (33 percent) of this 1995 to 2000 migrant population. Not surprisingly, the majority (54 percent) of movers remained within their original county of residence (U.S. Census Bureau, 2003); whereas 46 percent of movers relocated to a different county or state. Other U.S. Census Bureau estimates indicate that roughly 16 to 18 percent of households will change their address during a one-year period (based upon historical estimates from 1990 to 2000), with roughly one-half of these moves occurring within the same county (U.S. Census Bureau, 2001). As a verification of these U.S. Census Bureau estimates, recent longitudinal studies have found the percentage of movers to be similar to those reported above. For example, Brick et al. (1997) found 23 percent of their participants selected for follow-up had changed their telephone number (a surrogate for moving) in a one-year period. Booth and Johnson (1985) found that 41 percent of their initial participants had changed telephone numbers over a three-year period. And in Bogalusa, 8 percent of the study participants moved within six months of the study initiation, while 32 percent had moved after seven years (Webber et al., 1987). Considering these estimates of population mobility, it is reasonable to assume that over the course of an extended longitudinal study a significant percentage of the initial participants can be expected to relocate.

Thus, in a geographically clustered longitudinal cohort study like the NCS, one significant concern is the impact of population mobility (i.e., movement of study subjects from one geographic region to another) on the financial, logistical, and scientific objectives of the study. For example, there may be increased costs associated with data collection for participants that move a significant distance away from their original geographic area, but who do not move into another geographic area that is included in the NCS sample (e.g., to collect an environmental measure, it might be necessary to have traveling field data collection teams specifically to capture information for each family that moved). Additionally, there may be significant costs associated with tracking these individuals to their new residence and there may be implementation difficulties, such as standardization of measures, associated with the data collection activities for these mobile study subjects (in particular for those that move to a new location that is not in close proximity to any of the NCS data collection centers). Alternatively, if additional resources are unavailable, or if the increased resources needed to track and collect data on study subjects that relocate are determined to be too costly in relation to the information gained (e.g., if the subject has already completed most of the study or has already provided the majority of the necessary data), it may be preferable in certain cases to employ alternative data collection methods that would help to mitigate these additional costs (e.g., telephone interview in place of an in-person interview, mailing samplers rather than technician visits, etc.), or in some extreme cases to simply allow the subject to drop from the study. In this case, there would likely be scientific impacts in that mobility may increase study subject attrition or decrease the available information resulting in smaller sample sizes and/or less accurate data for statistical analyses of interest in the NCS.

In this paper we estimate the extent of subject mobility for the NCS as it pertains to a sample design in which the cohort is initially clustered into a finite number (e.g., between 50 and 300) of geographic regions represented by counties. In particular, given a set of initial geographic areas in the clustered design and using existing data sources that describe county-to-county flows of the US population, we construct transition matrices that allow projection of the number of NCS participants who are likely to move away from these original areas at different stages over the course of the study. In other words, given an initial cohort that is geographically clustered, we predict the geographic spread of that cohort over time. By investigating this geographic spread as a function of the number of clusters and for a number of different clustered designs, the results provide an indication of the potential impact of mobility on the NCS cohort. These projections can then be utilized in estimating the financial and scientific costs associated with subjects who move.

The remainder of this report is organized in the following manner. Section 2 presents the data utilized in estimating county-to-county move probabilities and briefly outlines the methods used in projecting the geographic dispersion of the initial cohort. Section 3 presents the results of applying these methods to a number of initial designs and illustrates the extent of geographic dispersion as a function of time and the number of geographic regions in the initial design. Finally, Section 4 discusses the relevant conclusions that result from this investigation.

2. Data and Methods

Based on the Census 2000 long form, the Census Bureau maintains a data source that indicates the county-to-county migration flows between the years 1995 and 2000 for the entire US population (http://www.census.gov/population/www/cen2000/ctytoctyflow.html). For every recorded combination of a migration from one county (county X) to another county (county Y), these data record the total number of individuals that resided in county X in 1995 and resided in county Y in 2000 (i.e., the migration flow from county X to county Y). Additionally, a county-specific gross migration data file that specifies the total population 5 years and over at year 2000, the portion of that population that were non-movers over the past 5 years, the portion of that population that were movers over the past 5 years, and the portion of that population that moved but remained in the same county over the past 5 years, is also available from the Census website (see http://www.census.gov/population/www/cen2000/phc-t22.html). By appropriately combining these two data sources we construct a 5-year transition probability matrix which identifies the probability of moving from county X to county Y over a five-year period for all pairwise combinations of X and Y (including where X=Y). For example, to compute the probability of moving from County X to County Y, we divide the proportion of individuals that moved from County X to County Y by the total number of individuals living in County X in 1995. This total number of individuals for 1995 is computed as the number of individuals that lived in County X in 1995 and remained there in 2000 plus the number of individuals that lived in County X in 1995 and moved to any other county within the United States in 2000. Several important assumptions that are incorporated in the construction of this transition matrix are as follows:

- Since the Census data does not provide information on the number of movers that relocate outside of the United States and does not provide information on the number of individuals living in a given county in 1995 that were deceased in 2000, the probabilistic transition matrix does not allow for transitions from living in a given county to dieing or to living outside the United States. However, these probabilities are likely very small especially for individuals that are age appropriate for the NCS, and thus should have very little impact on these results.
- A 1-year transition probability matrix could also be constructed by assuming that county-to-county flows are evenly distributed over the five-year interval and appropriately adjusting the number of individuals that remain in their current residence (so that the transition probabilities still sum to one). However, in order to remain consistent with the form of the Census data, we focus our results obtained using the 5-year transition probability matrix.
- The transition probability matrix is constructed based on the mobility of the entire population of the United States. However, in the NCS, the rate of mobility may be slightly higher since the population of interest in the NCS generally corresponds to the young adult (25 to 39) period of life, and the Census estimates suggest that these types of individuals may experience slightly higher mobility rates than the general population.

Using the constructed 5-year probability transition matrix and given the geographic representation of the initial cohort (i.e., the number of subjects in each selected county), we can project the geographic dispersion of the cohort in five-year increments by simply applying the transition probabilities to compute the expected number of people moving (or staying) from each of the current counties with a positive number of cohort members (i.e., year Y counties) into each county five years in the future. In other words, we simply compute the average dispersal of the cohort population across the counties of the United States. [Note that instead of estimating mobility via simulation, we instead compute the expected number of individuals in each county so that the average impact of subject mobility can be evaluated.]

As an example, Figure 1 displays the expected number of individuals in each county as a function of the year of the study. The upper left panel of the figure displays the counties represented in the initial, or year 1, cohort (in this case 51 counties selected with probability proportional to size sampling) by highlighting those counties in red. The other three panels of the figure continue to display these original counties highlighted in red, but also indicate the manner in which the cohort population has migrated to other counties by displaying different colors for each county in terms of the expected number of cohort members in that county. This geographic dispersion of the cohort population is displayed for year 5, year 10, and year 15 of the study and provides a graphical illustration of the dispersal of the NCS population across the United States for a design that initiates with 51 counties. In the analyses that follow, we also investigate designs that initiate with 100 and 250 clusters/counties in order to determine the extent to which designs with a larger number of initial clusters help minimize the negative impact of population mobility.

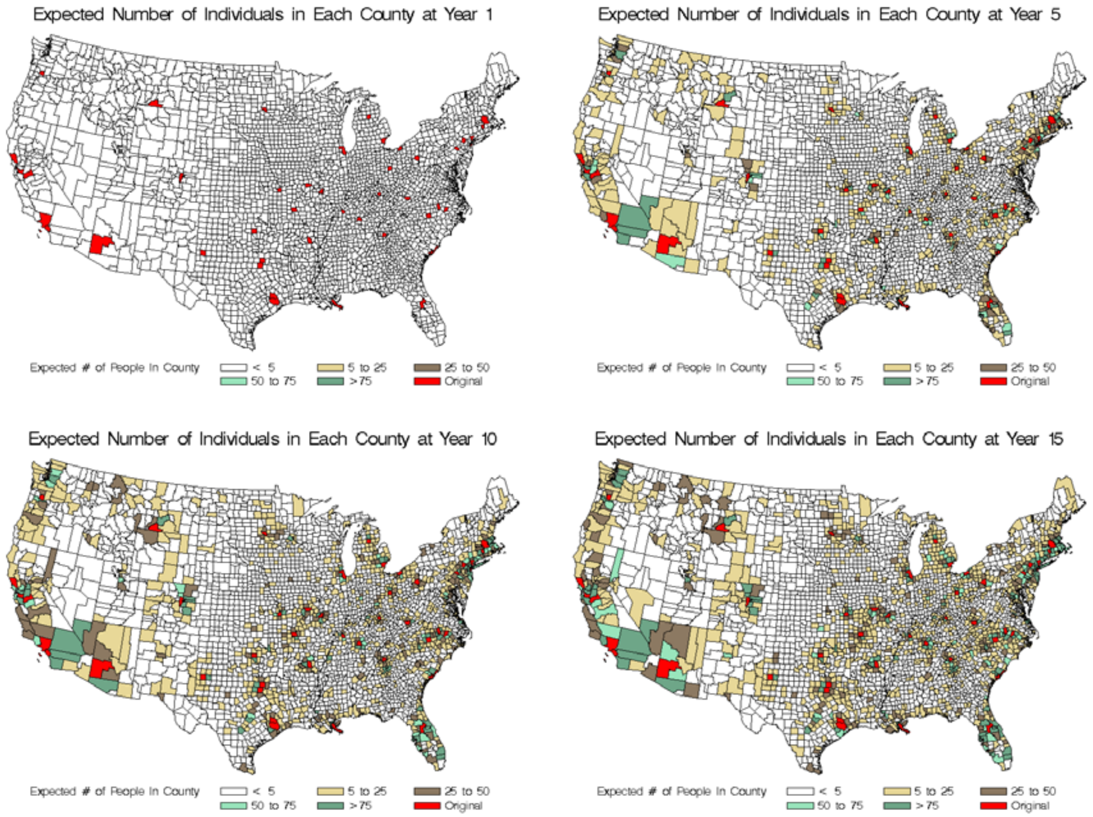
Expected Number of Individuals in Each County at Year 1
Expected Number of Individuals in Each County at Year 5
Expected Number of Individuals in Each County at Year 10
Expected Number of Individuals in Each County at Year 15
Expected # of People In County
< 5
5 to 25
25 to 50
50 to 75
>75
Original

Figure 1. Example illustration of the geographic dispersal of an initial cohort clustered into 51 counties.

In addition to the graphical illustration, and perhaps more importantly, these projections can also be used to compute statistics that measure the amount of mobility expected over time. For example, under the assumption that it will be more efficient to collect data in the originally selected counties (e.g., if these counties correspond to some sort of local data collection center), we may wish to assess the number of subjects for which it is necessary to implement alternative data collection procedures, such as alternative methods of collecting the desired data (e.g., a mobile data collection unit) or alternative types of data collection (e.g., increased phone interview data as opposed to biological or environmental data sampled at the home). This may be best estimated by projecting the number of subjects that are expected to move away from the original counties or to move to counties that are a significant distance (e.g., greater than 25, 50, or 100 miles) away from the original counties. [Note that county-to-county distances are calculated using the longitude and latitude of the centroids of the two candidate counties.] Additionally, it may be of interest to estimate the number of counties and the number of people living in counties that are a significant distance from the original counties and that have greater than 10 or 25 study subjects (or some other relatively large cluster of study subjects). This type of information might indicate the number of counties for which it might be reasonable to organize/send a mobile data collection unit (i.e., since a significant number of subjects reside in the county it may be cost-effective to send a mobile data collection unit to that county) and the number of people that would be "covered" by such a procedure.

As an example, Table 1 illustrates these types of summary statistics as a function of time for the geographic design depicted in Figure 1. The table displays the number of study subjects residing in the 51 original counties and the number of study subjects residing in counties that are greater than 25 and 50 miles from these original counties (in the results of Section 3 we also include these statistics for counties that are greater than 100 miles from the original counties), and the number of these "remote" counties, along with the number of study subjects residing in them, that contain greater than 10 subjects and greater than 25 subjects. It should be noted that the counties greater than 50 miles from the original counties are a subset of the counties that are greater than 25 miles from the original counties. Since the probabilistic transition matrix for county-to-county moves represents 5-year moving flows, these statistics are displayed at 5-year intervals beginning in Year 1 of the study and continuing to Year 20 of the study. In the results of Section 3 we provide similar tables for each of the geographically clustered designs that are evaluated.

Table 1. Impact of mobility for trial sample consisting of 51 original counties (geographic regions).

Year	# Subjects Living in Original Counties	# of Subjects in Counties w/in 25 Miles of Original Counties	# of Subjects Residing in Counties >25 Miles from Original Counties			# of Subjects Residing in Counties >50 Miles from Original Counties		
			All Counties [a]	Counties with >10 Subjects (# Counties)[b]	Counties with >25 Subjects (# Counties) [b]	All Counties [a]	Counties with >10 Subjects (# Counties) [b]	Counties with >25 Subjects (# Counties)[b]
1	100,000	0	0	0 (0)	0 (0)	0	0 (0)	0 (0)
5	83,831	2,915	13,254	8,949 (320)	5,709 (119)	9,238	5,565 (209)	3,369 (72)
10	71,047	4,920	24,033	18,716 (526)	14,388 (249)	17,018	12,332 (373)	9,020 (161)
15	60,872	6,273	32,855	27,286 (701)	21,352 (321)	23,610	18,623 (521)	13,826 (214)
20	52,720	7,162	40,118	34,226 (816)	27,374 (384)	29,226	23,869 (618)	18,428 (271)

[a] Number of subjects residing in the indicated counties (e.g., counties greater than 25 miles from original counties).

[b] Number of subjects residing in counties that are "far" away from the original counties and that contain greater than the indicated number of study subjects (e.g., counties greater than 25 miles from original counties that have greater than 25 subjects).

Of course, the geographically clustered design utilized and illustrated in Figure 1 and Table 1 represents only one realization of a candidate NCS design. Depending on the mobility associated with the selected counties and the number of selected counties, other candidate NCS designs may undergo a larger or smaller impact of subject mobility. However, since the number and location of the geographic regions corresponding to the initial cohort have not yet been selected, we evaluate the impact of mobility for several different designs. For each design we assume a total cohort size of 100,000 individuals and we assume that each of the geographic regions (counties) will contain the same number of participants (e.g., 2000 subjects in each of 50 counties or 1000 subjects in each of 100 counties). Thus, in order to specify a geographically clustered design we need only identify the number and location of the geographic regions selected in the design. We also assume no attrition over time for simplicity.

To investigate the geographic mobility as a function of the number of geographic regions in the initial design, we evaluate designs with approximately 50, 100, and 250 counties represented in the initial cohort. To select the actual locations (counties) corresponding to these designs two approaches are utilized. First, we evaluate designs in which the counties are selected with probability proportional to size (population size) sampling from eight strata consisting of four regions of the United States and rural versus urban counties (see Strauss et al., 2004 for more details). Since this probabilistic sampling of counties will result in a different set of initial counties for each sample realization, we evaluate the impact of mobility for 50 realizations of each type of design and average the results over these 50 realizations. Second, we evaluate three geographically clustered designs (with 50, 100, and 250 regions) that are selected to coincide with many of the largest (in population) counties across the continental United States while maintaining the restriction of selecting at least one county in each state. These designs select the largest county in each of the 48 continental states (resulting in 48 initial counties), and then select the rest of the needed counties by choosing the largest remaining counties across the U.S. Since these designs are selected specifically to correspond to the largest portion of the U.S. population while also enforcing geographic diversity in the initial sampling stage that selects one county in each state, evaluation of them may provide an indication of a “best-case” scenario in terms of cohort mobility (i.e., designs that have the largest portion of the cohort residing in one of the initial counties as a function of time).

3. Results

Table 2 displays the summary statistics corresponding to the case where counties are selected with probability proportional to size sampling from eight different strata, and Figures 2 and 3 provide a graphical display of these values as a function of time. As described above, for each of the three types of designs (50 counties, 100 counties, 250 counties), 50 realizations are evaluated and the average cohort mobility across these 50 realizations is displayed in the table and figures. As suggested by the Census Bureau estimates discussed previously, these results indicate that a significant portion of the cohort will move away from the initial set of counties. For example, at Year 5 of the study approximately 16,500 study subjects will move away from the initial set of counties in a 50-county design, approximately 15,000 subjects will move away from the initial set of counties in a 100-county design, and approximately 11,500 subjects will move away from the initial set of counties in a 250-county design. However, if the catchment area for each of the initial data collection centers included all counties within a 100-mile radius (as measured by the distance between county centroids), the expected number of study subjects who move away from these catchment areas by year 5 is much less (4600 for a 50 cluster design, 2500 for a 100 cluster design, and 800 for a 250 cluster design).

Additionally, we can project the number of counties that are outside the catchment areas (25, 50, and 100 mile radius) of the initial counties and that contain enough study subjects (e.g., 10 or 25) to make it cost effective for the NCS to consider sending out a mobile data collection center (similar to the trailers used in NHANES). Table 2 suggests that at Year 5, 97 percent of the original study population could be accessed with a 50 cluster design – with approximately 83,500 subjects residing directly in one of the original data collection center counties, 12,000 subjects living outside the original counties but within a 100-mile radius, and 1,500 subjects in 32 counties with 25 or more subjects that would be visited by a mobile data collection center. Similarly, by Year 15 this approach would cover 94.5 percent of the original study population with approximately 60,000 subjects residing in one of the original counties, 28,000 subjects living outside the original counties but within a 100-mile radius, and 6,500 subjects living in 103 counties with 25 or more subjects that would be visited by the mobile data collection center.

Table 2. Impact of mobility for geographically clustered designs in which counties are selected probabilistically.

# of Original Counties	Year	# Subjects Living in Original Counties	# of Subjects in Counties w/in 25 Miles of Original Counties	# of Subjects Residing in Counties >25 Miles from Original Counties			# of Subjects Residing in Counties >50 Miles from Original Counties			# of Subjects Residing in Counties >100 Miles from Original Counties		
				All Counties [a]	Counties with >10 Subjects (# Counties)[b]	Counties with >25 Subjects (# Counties) [b]	All Counties [a]	Counties with >10 Subjects (# Counties) [b]	Counties with >25 Subjects (# Counties)[b]	All Counties[a]	Counties with >10 Subjects (# Counties) [b]	Counties with >25 Subjects (# Counties) [b]
51	1	100,000	0	0	0 (0)	0 (0)	0	0 (0)	0 (0)	0	0 (0)	0 (0)
	5	83,445	3,349	13,206	9,030 (316)	5,825 (114)	8,553	5,108 (196)	3,018 (63)	4,590	2,607 (103)	1,494 (32)
	10	70,353	5,682	23,965	18,776 (519)	14,319 (235)	15,785	11,232 (346)	8,048 (141)	8,520	5,803 (183)	4,115 (74)
	15	59,927	7,282	32,791	27,051 (666)	21,503 (313)	21,942	16,765 (465)	12,594 (197)	11,912	8,719 (246)	6,495 (103)
	20	51,567	8,356	40,077	33,929 (776)	27,625 (377)	27,215	21,584 (558)	16,685 (244)	14,857	11,299 (297)	8,677 (129)
100	1	100,000	0	0	0 (0)	0 (0)	0	0 (0)	0 (0)	0	0 (0)	0 (0)
	5	85,056	3,705	11,239	6,954 (280)	3,986 (88)	6,300	3,271 (143)	1,671 (39)	2,493	1,177 (54)	559 (13)
	10	73,250	6,363	20,387	15,024 (478)	10,576 (195)	11,606	7,505 (264)	4,916 (97)	4,609	2,775 (102)	1,744 (36)
	15	63,856	8,257	27,887	21,951 (621)	16,485 (273)	16,109	11,352 (357)	7,987 (141)	6,421	4,257 (140)	2,901 (53)
	20	56,330	9,594	34,076	27,788 (733)	21,496 (333)	19,955	14,786 (438)	10,738 (177)	7,982	5,582 (172)	3,966 (68)
250	1	100,000	0	0	0 (0)	0 (0)	0	0 (0)	0 (0)	0	0 (0)	0 (0)
	5	88,500	3,713	7,787	3,588 (183)	1,433 (40)	3,291	1,213 (67)	365 (11)	835	252 (14)	79 (2)
	10	79,457	6,492	14,051	8,754 (365)	4,888 (113)	5,997	3,097 (138)	1,586 (40)	1,509	709 (34)	321 (8)
	15	72,292	8,572	19,137	13,303 (498)	8,251 (172)	8,248	4,858 (196)	2,827 (64)	2,064	1,126 (48)	578 (13)
	20	66,575	10,129	23,296	17,149 (600)	11,204 (220)	10,135	6,428 (245)	3,939 (83)	2,525	1,500 (60)	859 (19)

[a] Number of subjects residing in the indicated counties (e.g., counties greater than 25 miles from original counties).

[b] Number of subjects residing in counties that are “far” away from the original counties and that contain greater than the indicated number of study subjects (e.g., counties greater than 25 miles from original counties that have greater than 25 subjects).

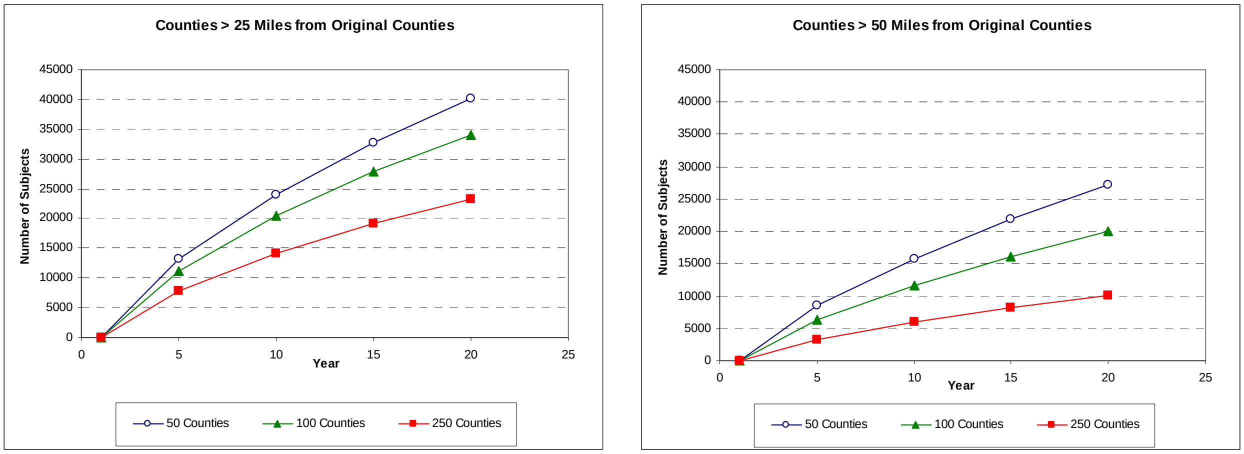

Figure 2. **Number of subjects living in counties that are greater than 25 miles from the original counties (left panel) and greater than 50 miles from the original counties (right panel) as a function of time and the number of counties selected in the original design (probabilistic selection of counties).**

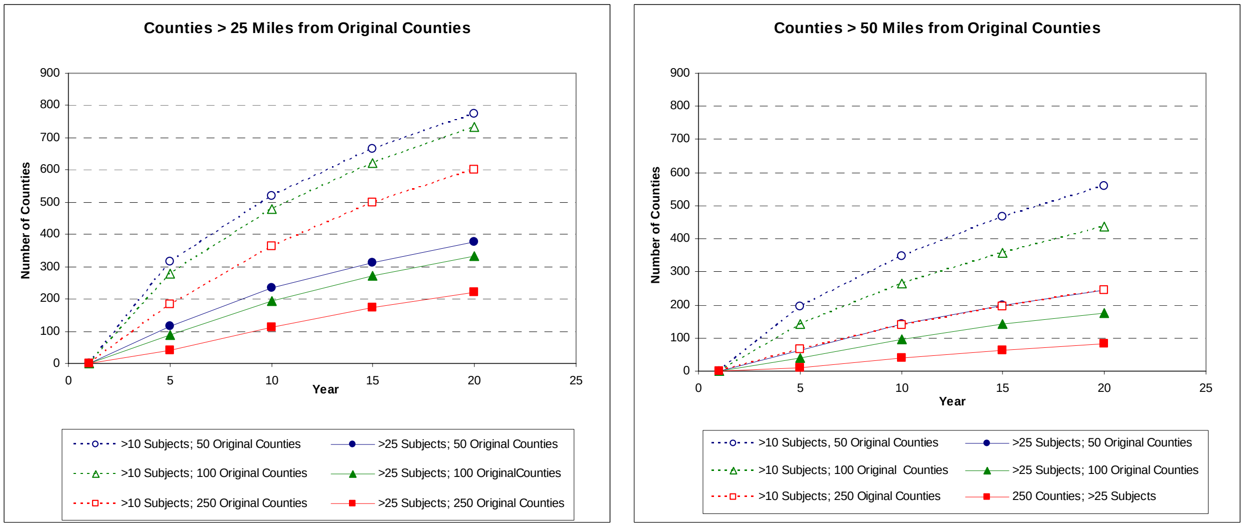

Figure 3. **Number of counties that have at least 10 (or 25) study subjects and that are greater than 25 (left panel) or 50 (right panel) miles away from the original counties as a function of time (probabilistic selection of counties).**

Similar to Table 2 and Figures 2 and 3, Table 3 and Figures 4 and 5 display the impact of mobility for designs in which the counties were purposively selected to correspond to the largest (in population) counties across the United States. In so doing, these designs may represent a “best-case” scenario in terms of cohort mobility. Comparing the numbers in this table to the corresponding numbers displayed in Table 2 there appears to be a slightly smaller amount of cohort mobility in this case (as expected); however, the differences are not dramatic suggesting that there is only a small impact of the manner in which the initial set of counties are selected.

Table 3. Impact of mobility for geographically clustered designs in which the largest county in each state is selected and the remaining counties are selected to correspond to the next largest counties (by population size).

# of Original Counties	Year	# Subjects Living in Original Counties	# of Subjects in Counties w/in 25 Miles of Original Counties	# of Subjects Residing in Counties >25 Miles from Original Counties			# of Subjects Residing in Counties >50 Miles from Original Counties			# of Subjects Residing in Counties >100 Miles from Original Counties		
				All Counties [a]	Counties with >10 Subjects (# Counties) [b]	Counties with >25 Subjects (# Counties) [b]	All Counties [a]	Counties with >10 Subjects (# Counties) [b]	Counties with >25 Subjects (# Counties) [b]	All Counties[a]	Counties with >10 Subjects (# Counties) [b]	Counties with >25 Subjects (# Counties) [b]
50	1	100,000	0	0	0 (0)	0 (0)	0	0 (0)	0 (0)	0	0 (0)	0 (0)
	5	83,642	3,682	12,676	8,299 (292)	5,275 (101)	7,991	4,482 (179)	2,519 (52)	4,944	2,906 (120)	1,593 (35)
	10	70,871	6,203	22,926	17,388 (493)	13,093 (218)	14,731	9,949 (316)	6,979 (128)	9,143	6,440 (205)	4,600 (89)
	15	60,816	7,913	31,271	25,093 (643)	19,425 (285)	20,451	15,014 (438)	10,999 (179)	12,729	9,600 (270)	7,326 (124)
	20	52,838	9,053	38,109	31,389 (748)	25,239 (357)	25,331	19,317 (524)	14,760 (228)	15,808	12,323 (318)	9,718 (151)
100	1	100,000	0	0	0 (0)	0 (0)	0	0 (0)	0 (0)	0	0 (0)	0 (0)
	5	86,056	3,534	10,410	5,885 (249)	3,150 (71)	6,250	3,093 (142)	1,497 (36)	3,113	1,767 (81)	851 (20)
	10	75,098	6,039	18,863	13,132 (451)	8,810 (173)	11,488	7,198 (263)	4,642 (98)	5,735	3,915 (133)	2,743 (58)
	15	66,418	7,801	25,781	19,569 (611)	13,899 (245)	15,909	11,013 (365)	7,570 (141)	7,958	5,950 (181)	4,310 (76)
	20	59,490	9,029	31,481	24,639 (702)	18,223 (300)	19,662	14,144 (428)	10,200 (176)	9,853	7,604 (208)	5,778 (93)
250	1	100,000	0	0	0 (0)	0 (0)	0	0 (0)	0 (0)	0	0 (0)	0 (0)
	5	90,219	2,549	7,231	2,865 (161)	915 (26)	3,873	1,366 (79)	373 (11)	1,029	377 (23)	72 (2)
	10	82,454	4,457	13,089	7,298 (327)	3,519 (86)	7,045	3,562 (165)	1,705 (44)	1,888	991 (47)	445 (12)
	15	76,235	5,884	17,881	11,546 (473)	6,638 (152)	9,674	5,741 (243)	3,183 (74)	2,609	1,583 (66)	936 (23)
	20	71,212	6,954	21,834	15,085 (576)	9,161 (193)	11,873	7,586 (300)	4,401 (93)	3,220	2,089 (80)	1,255 (27)

[a] Number of subjects residing in the indicated counties (e.g., counties greater than 25 miles from original counties).

[b] Number of subjects residing in counties that are “far” away from the original counties and that contain greater than the indicated number of study subjects (e.g., counties greater than 25 miles from original counties that have greater than 25 subjects).

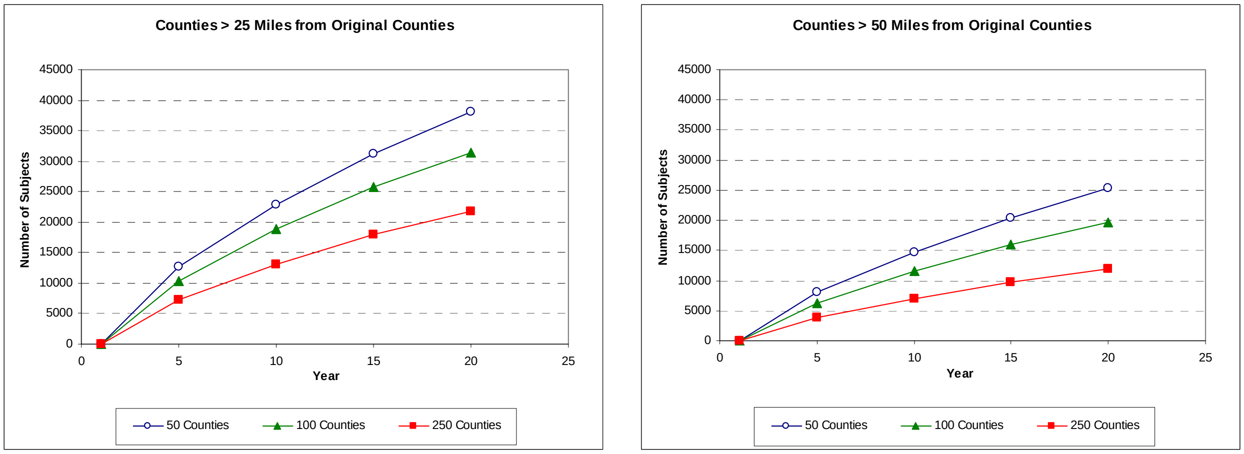

Figure 4. **Number of subjects living in counties that are greater than 25 miles from the original counties (left panel) and greater than 50 miles from the original counties (right panel) as a function of time and the number of counties selected in the original design (counties selected based on population size).**

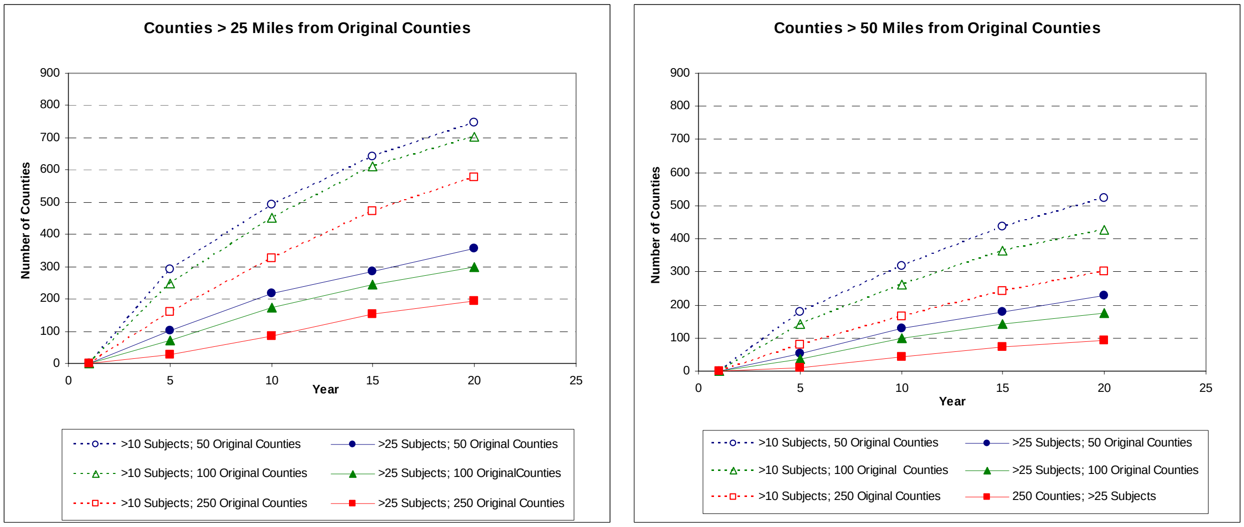

Figure 5. **Number of counties that have at least 10 (or 25) study subjects and that are greater than 25 (left panel) or 50 (right panel) miles away from the original counties as a function of time (counties selected based on population size).**

4. Conclusions

The results presented above demonstrate the potential impact of mobility on the NCS cohort by estimating the number of study subjects that reside outside of the original set of counties represented in the cohort and/or the portion of the cohort that resides a significant distance away from these counties (e.g., resides in counties greater than 25 miles away from the original set of counties). For example, the estimates provided in Table 2 suggest that if a 50-

county design is utilized approximately 16,500 study subjects will reside outside of those counties at Year 5 of the study, 30,000 at Year 10, and 40,000 at Year 15. Increasing the size of the catchment area for each of the original 50 counties, we estimate that approximately 4,500 study subjects will reside in counties that are greater than 100 miles from the original 50 counties at Year 5, 8,500 at Year 10, and 12,000 at Year 15. Alternatively, if a 250-county design is utilized these numbers become slightly smaller with approximately 11,500, 20,500, and 28,000 study subjects living outside of the original counties at Years 5, 10, and 15, respectively, and approximately 1,000, 1,500, and 2,000 subjects residing in counties that are greater than 100 miles from the original 250 counties at Years 5, 10, and 15, respectively.

This issue of the number of geographic regions that should be selected in the NCS has been discussed in a number of contexts related to statistical efficiency, implementation difficulties, cost efficiency, etc. In terms of mobility, not surprisingly, if more regions are selected in the initial design then the study will have greater coverage of the mobile cohort over time. In other words, if cohort mobility were the only consideration for choosing a design, then designs with fewer geographic regions would be less optimal. Taking into consideration factors such as the financial cost of tracking and collecting data from study subjects that move as well as the costs of starting and maintaining data collection operations in each of the initially selected counties, the optimal design becomes less apparent. Certainly, designs with fewer geographic regions initially would likely experience a larger financial expense to track and collect data from study subjects that move. However, designs that initiate with a larger number of geographic regions would likely experience a larger amount of up-front and maintenance costs for organizing data collection components in the larger number of regions.

To compare the 50, 100, and 250-county designs in a more concrete manner, suppose that an established data collection center would be able to track and collect data from those study subjects that resided within a 100-mile radius. Additionally, suppose that we have a nominal goal of covering 95 percent of the original cohort with local or mobile data collection centers (i.e., less than 5,000 study subjects can reside in a county outside of the 100-mile catchment area of a data collection center or not covered by a mobile collection unit). As displayed in Table 2, for a 250-County design there would be no need for mobile data collection centers at any time over the course of the study since the portion of the population that resides in counties that are greater than 100 miles away from the original counties is always less than 5,000 individuals. On the other hand, for the 50- and 100-County designs it would be necessary to establish mobile data collection centers over the course of the study in order to maintain coverage of at least 95,000 study subjects. In particular, for the 100-County design it would be necessary to establish mobile data collection centers beginning prior to year 15 of the study, and for the 50-County design, mobile data collection centers would need to be established prior to year 10 of the study. Assuming that the mobile units can cover a number of regions (e.g., 10 to 20) in a year, it would appear that for both the 50 and 100-County designs a large portion of the cohort can be covered with the data collection centers in the original counties plus a small number of additional mobile data collection centers (e.g., judging by the number of counties that are greater than 100 miles from the original counties but have a relatively large number of study subjects). Additionally, it is clear that much of the work for the mobile collection centers would only be necessary in the later years of the study.

This simple example suggests that designs initiating with on the order of 50 to 100 counties, each having a data collection center, and adding mobile data collection centers over time in order to maintain coverage of at least 95 percent of the cohort population, may be more cost-efficient than a design that initiates with 250 counties since they would likely require a much smaller number of data collection centers over the course of the study. A useful follow-on analysis to these results would be to provide an accurate assessment of the number of regions that would need to be covered by mobile centers in order for the 50- and 100-county designs to maintain coverage of at least 95 percent (or some other significant proportion) of the cohort. Presumably, by strategically locating these mobile centers, coverage of the desired portion of the cohort population could be accomplished in an efficient manner. In any case, based on these analyses there does not appear to be a strong argument that a 250-county design would be more optimal in terms of cohort mobility. In fact, it appears that a 50- or 100-county design, with additional mobile collection centers added over the course of the study, would be more cost-efficient.

More generally, these results demonstrate that for the geographically clustered designs evaluated here (and likely for any other plausible approaches to selecting such a design) a significant portion of the cohort will move out of the original set of selected counties over the course of the study (and/or will change residences within the same county). Thus, if maintaining contact and study enrollment of subjects that move is an important component of the NCS, then there should be procedures and tracking mechanisms to maintain contact with and to collect data from these subjects (in particular for subjects that move away from the region in which they were originally recruited). For example, one reasonable mechanism may be to refer subjects that move away from one data collection center but within the catchment area of an already established data collection center to those centers for the necessary data collection. Alternatively, all study subjects that move into regions that are not covered by an established collection center might be referred to a single organization capable of following these individuals across the nation (or a mobile data collection center could be utilized if a certain region appears to have a significant portion of the cohort population).

Tracking of participants over the course of the study is a different type of activity which may involve tasks such as 1) annual (or more frequent) verification of addresses and contact information via telephone or postcard mailings, 2) searching for individuals/families via public databases when they are no longer at their previous address and have left no forwarding contact information, 3) notifying the data collection coordinating organizations of participants moving into and out of their areas, and 4) developing and providing regular newsletters and/or other mailings (e.g., birthday cards) to participants.

Of course, as with any study component, the costs of tracking and collecting data from participants who moved should be weighed against the overall benefits of continuing to follow that subject, the investment already incurred by the NCS for that family, and the timing of when the child leaves the study because of a move. Some of the salient issues when determining whether to follow study subjects that move are:

- Are the health outcomes of interest rare or require many years for diagnosis?
- Do movers represent a “rare” population of interest?

- Are the characteristics of movers generally different than those of non-movers so that there is a plausible relationship between mobility and health outcome (i.e., is movement by subjects related to the likelihood of exposure and subsequent health effects)?
- Are there significant changes in exposure as a result of moving?
- Are respondents that move still representative of the population of interest?
- Can tracing or locating participants that move be completed within reasonable costs?
- Can data collection from participants who move be completed within reasonable costs?
- Do alternative methods for data collection need to be considered for participants that move?
- How does moving impact the timing or schedule of data collection?

For the NCS, affirmative responses to the first five questions would suggest that movers should be followed and data collection should be continued from these participants. For example, if environmental exposure during early childhood does not result in an adverse health outcome for many months, years, or even decades, failure to follow children who were initial study participants, but subsequently moved, may introduce bias into the estimates of the relationship between exposure and outcome. Answers to the remaining questions still need to be ascertained, and may vary from respondent to respondent and by data collection method. Certainly, study subjects that move within the same county and/or move to a location that is in close proximity to an already established data collection center should be followed since the additional costs of tracking and data collection for these subjects is likely minimal. However, for study participants that move to counties or geographic regions that are far from any of the established data collection centers, the costs of certain types of data collection, such as environmental samples, may be significantly larger since it may be necessary for trained personnel to travel to those locations in order to collect the necessary data. Perhaps it would be acceptable for these subjects to move to a less detailed data collection protocol, such as a phone-based interview, in order to maintain these subjects in the cohort while minimizing the costs associated with collecting their data.

Finally, it should be noted that the methods utilized in this report to project the geographic dispersal of a candidate NCS cohort should be re-applied once the actual NCS cohort (or the geographic regions of the NCS cohort) has been selected. This would allow a more accurate portrayal of the expected cohort mobility for the selected cohort, and provide study leaders important information that could be useful in planning where, when, and to what degree the cohort will disperse across the United States. Additionally, 1-year mobility projections can be evaluated to assist in study planning.

5. References

Webber L, Frank G, Smoak C, Freedman D, and Berenson G. 1987. Cardiovascular Risk Factors From Birth to 7 Years of Age: The Bogalusa Heart Study, Design and Participation. Pediatrics 80: 767-78.

Booth A, Johnson D. 1985. Tracking Respondents in a Telephone interview Panel Selected by Random Digit Dialing. Sociological Methods and Research. 14:53-64.

Brick M, Collins M, Davies E, Chandler K. 1997. Feasibility of Conducting Follow-up Surveys in the National Household Education Survey. U.S. Department of Education. National Center for Education Statistics, NCES 97-335. Washington, DC.

Strauss W, Lehman J, Menkedick J, Ryan L, Pivetz T, McMillan N, Pierce B, Rust S. 2004. Evaluation of Sampling Design Options for the National Children's Study. White paper prepared for the National Children's Study Program Office, March 19, 2004.

U.S. Census Bureau. 2003. Migration of the Young, Single, and College Educated: 1995 to 2000. Census 2000 Special Reports.

U.S. Census Bureau. 2001. Geographical Mobility Population Characteristics March 1999 to March 2000. Washington, DC.