Life Course Socioeconomic Status (LC-SES): Manual of Procedures and Overview of Data Sets Available

Table of Contents

1. Introduction

Life Course SES, Social Context and Cardiovascular Disease (LC-SES) is a four-year study sponsored by the National Heart, Lung, and Blood Institute. Its purpose is to study the explanatory mechanisms for the association of historical and contemporary socio-economic status (SES) with cardiovascular disease (CVD) in a large, population-based bi-ethnic cohort of men and women. LC-SES is an ancillary study to the Atherosclerosis Risk in Communities (ARIC) study.

Between 2001 and 2002, survivors from the Atheroclerosis Risk in Communities (ARIC) Study were queried about their childhood and earlier adulthood socio-environmental circumstances. Information collected included childhood and early adulthood individual-level measures of SES as well as childhood and early adulthood place of residence. Place of residence was linked with census level measures of the social environment from the appropriate time period. The childhood and early adulthood SES information collected as part of the LC-SES ARIC Ancillary Study is being used with the extant individual and census tract level SES data collected at the ARIC baseline exam and the subsequent ARIC examinations to construct profiles of SES across the life course. This is being used in conjunction with the extant behavioral and biomedical risk factor data collected over the course of the ARIC study to attempt to elucidate pathways linking SES / the social environment to cardiovascular disease (CVD) across the life course.

Most efforts to date to study the influence of the social context on health have focused only on the current context. Thus, a unique aspect of the LC-SES study is the collection of historical residential information including linking this information to historical census-based measures of the social environment. However, linking self-reported previous places of residence with historical census data is more complex than linking current address with more recent census data. As established procedures for doing this were not identified in the literature, a major purpose of this manual is to provide an overview and documentation of the procedures developed during the LC-SES study.

2. Data Sources

There are two primary sources of data for the LC-SES Ancillary Study: Individual-level SES information, collected in the LC-SES ARIC Annual Follow-up telephone interview and as part of ARIC Study (Visits 1-4); and neighborhood SES indicators, where the census tract(or county) is the unit of observation and pertinent decennial censuses years are the source of the data.

2.1. LC-SES Study

The LC-SES questionnaire was administered between 2001and 2002, as part of the ARIC Annual telephone follow-up (AFU). Its official title is the "Residential / Occupational History Form (SES-B)". The purpose of the questionnaire was to obtain information on childhood and early adulthood socioeconomic circumstances as such information was not collected during the ARIC examinations. Childhood measures

included mother's and father's occupation and education, as well as home ownership and city and state of residence. Early adult measures included types and characteristics of occupations at ages 30, 40, and 50, address and home ownership at ages 30, 40, and 50, and information related to participation in the armed services. A copy of complete SES-B Questionnaire can be found in on this Website under *Forms & Manuals*.

2.2. ARIC Study

The data collected as part of the LC-SES Study represents only a portion of the SES data that will be used. Another source of data on earlier life SES is the SES-A questionnaire, which had been administered during the ARIC Visit 4 examination. It includes information about parents' education as well as own occupation between ages 25-44 and current occupation and family income. A copy of the SES-A questionnaire can be found in *Forms & Manuals* on this Website.

Additional individual-level SES and socio-demographic measures pertaining to mid to later life were collected during ARIC visits 1-4 as well as selected annual telephone follow-ups. These variables, as well as relevant items from the SES-A questionnaire, have been incorporated into the LC-SES dataset.

2.3. Census-Based Socioenvironmental Data

One of the goals of the LC-SES Study is to determine the extent to which current and historical neighborhood context modify the association of individual-level life course SES exposures with CVD-related events.

Historical addresses were geocoded at the level of the appropriate (1960, 1970 or 1980) census tract, allowing us to link census-based area-level (neighborhood) measures to individual records and to examine the impact of the long-term social environment on CVD-related health outcomes in adulthood.

Addresses from the ARIC visits 1-4 were geocoded and linked to 1990 census data. Addresses from visit 4 were also linked to 2000 census data, as the time period covered corresponded most closely to 2000.

3. Quality Assessment of Data Collected in LC-SES Study

The LC-SES investigators developed a series of steps to edit & check the SES-B data prior to the creation of the final analysis files. The general steps are outlined below:

3.1 Procedures for editing & processing data

Data collection took place over a period of approximately one year. The Collaborative Studies Coordinating Center at UNC (CSCC) received the SES-B telephone interview results from the four ARIC field centers at intervals over the data

collection period and released restricted-use partial datasets to the LC-SES investigators with encrypted IDs to protect participants' confidentiality. These intermediate datasets were used to develop programs for "cleaning" the addresses.

Occasionally this resulted in feedback to the ARIC interviewers, specifically if repeated errors were noticed (e.g., consistent misspelling of a city name).

After the final data retrieval there were a total of 12,716 observations. The final dataset released by the CSCC was an edited file with participants' addresses but encrypted ARIC study ID's. The historical addresses from the dataset were sent to a commercial geocoding firm after being run through our data "cleaning" programs. When the geocoding process was complete, data sets were prepared that included Federal Information Processing Standards (FIPS) codes corresponding to the geographical areas in which participants resided. These were returned to CSCC where personnel removed all address information and then added the participants' actual study IDs .

3.2. Defining skipped vs. missing data

The primary purpose of the quality control assessment was to evaluate the extent of missing data. A SAS program was written to differentiate true missing values from those missing as a result of skip patterns inherent in the questionnaire. Major reasons for the skip patterns included:
• Participants who attended visit 4 and responded to the SES-A questionnaire were not re-queried about their parents' education in the SES-B interview.
• All participants were asked about place of residence at age 30; participants were queried selectively about place of residence at ages 40 and 50, depending on their age at the ARIC baseline exam. For example, if a participant was between ages 50 and 59 years of age at the baseline examination, he/she was not queried about his/her address at age 50.
• Skip patterns were built into the SES-B questionnaire reflecting the logic of the questions. For example, if a participant answered 'yes' to Question 1, they were directed to skip Questions 2-4 and answer Question 5; if they answered 'no' they were asked Question 2a.

Once we were able to distinguish true 'missings' from those resulting from skip patterns, patterns of true missing data and 'unknown' values were examined**.**

4. Preparation of Addresses for Geocoding

Missing and inaccurate address information can substantially reduce the success of the geocode matching, thus, to maximize match rates, it is essential to edit residential addresses for obvious problems before submitting them to a commercial geocoding vendor.

4.1. ARIC Visit 1-4 (V1-V4) Addresses

In addition to reviewing and correcting the self-reported cities, counties and states for Visits 1-4, we reviewed street addresses to see if elements necessary for geocoding were present. If they were deemed adequate, the address was sent to the commercial geocoder. Major reasons that addresses were determined to be not adequate included:
• State missing (most often military addresses)
• All address fields missing
• C/O (in care of)
• Temporary address
• Address not in the U.S
• Apartment name without address or where no address could be located
• Only address is PO Box

Table 1 summarizes characteristics of the addresses by whether or not they were sent to the geocoder.

Table 1: Visit 1 – Visit 4 Addresses by Type of Information Provided

	Visit 1 (n=15792)	Visit 2 (n=14348)	Visit 3 (n=12885)	Visit 4 (n=11656)
SENT (N)		(14327)	(13887)	
Apparently complete address	14260	13769	12280	11166
Partial address	25	34	29	59
Zip code only	3	3		4
Street name only	125	18	32	7
PO box & address	0	0	72	0
Rural Route number	1052	209	101	74
TOTAL	15465	14033	12514	11310
NOT SENT (N)				
Address not adequate	13	16	10	7
PO Box only	314	278	363	339
TOTAL	327	294	373	346

4.2. Historical Addresses

The SES-B questionnaire queried participants' address at various life epochs: childhood and ages 30, 40, and 50 years. As a first step in preparing these addresses for further processing, the spellings of all cities, counties, and states were checked and, when possible, corrected. States were corrected for misspellings and standardized to the two digit abbreviations used by the US Postal Service. Within

each state, the cities and counties were hand-checked against look-up tables available through the internet (http://www.accesschecks.com/xref.htm). When a city was provided but no county listed we used either Map Quest (www.mapquest.com) or the National Association of Counties (www.naco.org) to identify the corresponding county, if extant. For the major study states (Maryland, Mississippi, Minnesota and North Carolina), we also verified that changes in counties had not occurred during the years of interest. When mistakes were found, they were corrected and 'flag variables' were created to track the number and types of errors. Flag variables were created for each set of address questions (childhood, age 30, age 40, and age 50).

4.2.1. Childhood Address Correction Summary

Fewer than 15% of the childhood addresses required any editing for the city, county, or state. Of the errors identified, more than 70% were misspellings. Misspellings and incorrect counties accounted for over 90% of the errors for childhood counties. Finally, nearly all of the edits to childhood state were misspellings or simple substitutions of the US Postal Code state abbreviation.

4.2.2. Age 30 Addresses

All 12,724 LC-SES participants were asked their residence at age 30 years. Approximately 80% of the reported cities, 85% of the reported counties, and 75% of the reported states appeared correct.

4.2.3. Age 40 Addresses

The 8,947 participants who were ages 50 and older at the time of the ARIC baseline examination were queried about their place of residence at age 40 years. Approximately 81% of the reported cities, 92% of the counties, and 78% of the states did not require editing.

4.2.4. Age 50 Addresses (sesb25a-sesb25e)

Participants who were ages 60 and older at the time of the ARIC baseline examination were asked to recall their place of residence at age 50 years (n=2483). Approximately 76% of the reported cities, 95% of the counties, and 80% of the states did not require editing.

5. Geocoding

The process of geocoding involves assigning geographic reference coordinates to non-geographic data. Addresses which include street names and house or structure number are linked to geographic databases which contain this same information. Geocoding software attempts to match the address from each observation in the non-geographic file with a street segment in the geographic database which includes street names, street addresses ranges on each side of the street, and street latitude/longitude coordinates. Commercial address geocoding services use data derived from the U.S. Census TIGER/Line® files.

Because the geographic extent and identifying codes for census units (tracts, etc.) are also stored in the TIGER/Line® files, the individual address information can be tagged with appropriate census codes. (Ref: 'Linking Survey Respondents' Residential Addresses to Contextual Data: Address Geocoding Issues, Udry, J.R., Carolina Population Center, 1995).

Census tracts were assigned to contemporary (Visit 1-4) and historical addresses. When commercial geocoding was not successful additional procedures were used to increase match rates.

5.1. Previous Geocoding Efforts in the ARIC Study

Visit 1 addresses, collected in 1987-89, were originally geocoded by a commercial geocoder to 1990 census boundaries, under the direction of one of the LC-SES investigators as part of a dissertation project. When the geocode was successful, an address was associated with a FIPs code, which included all census boundary information for that location (state, county, tract, block group). An unexpectedly high proportion of Washington County, MD. addresses did not geocode, apparently due to a change in addresses in the early 1990's as part of the state's effort to improve emergency response. Street numbers were changed to conform to a grid address system and street and road names were changed to remove rural routes and duplicated names. In order to minimize missing data for the participants with these addresses changes, the investigator solicited the help of Washington County ARIC field center personnel who were familiar with the area. With the use of a street map and 1990 census tract boundaries, they attempted to place the affected addresses into their correct geographic location so that tract numbers could be manually assigned.

Visit 3 addresses were originally geocoded in 1990 by a commercial company, Vendor A, to the 1990 census boundaries as part of an ARIC ancillary study.

5.2. Geocoding Vendor Selected and Vender Match Codes

For our work in the LC-SES study we chose a firm hereafter referred to as Vendor B. This decision was based on the recommendation of others who had systematically evaluated vendors as part of a geocoding accuracy study. Additional efforts aimed at evaluating the repeatability and accuracy of the geocodes provided by Vendor B over the course of our study are described later in section 5.5 of this manual.

Addresses sent to Vendor B were returned with geographic coordinates (latitude and longitude), a FIPS code, and a match code indicating the level of precision of each match. Match codes fall in three categories: A (address), A (zip code centroid), E (no match). For our purposes, we decided to accept only addresses which matched at least at the level of the census tract (as this will be the smallest geographical unit at which we will aggregate census data). This includes A codes and selected Z codes matching with tract level accuracy. *Table 2* illustrates the codes we considered acceptable**.**

Table 2: Definition of Match Codes Assigned by Vendor B

Code	Definition	Acceptable
ASn	Indicates a house range address geocode. This is the most accurate geocode available. The digit at the end indicates the following:	yes
AS0	Best location	yes
AS1	Street side is unknown. The Census FIPS Block ID from the left side is assigned, however no offset is assigned, the point is placed directly on the intersection.	yes
AS2	Address was interpolated on to a TIGER segment that did not contain address ranges initially.	yes
AS3	Both1and2	yes
AX3	Indicates an intersection geocode. The street side can not be determined, and no address ranges can be assigned, hence the ì3" is returned to be consistent with AS codes	yes
Zxxx	ZIP+4Centroid Match	
ZBxx	Indicates Block Group accuracy(most accurate)	yes
ZTxx	Indicates Census Tract accuracy	yes
ZCxx	Indicates unclassified Census accuracy. Normally accurate to at least the County level	yes
ZC7N	Location derived from a ZIP Code Centroid – unclassified Census accuracy – normally accurate to the county level –indicates a location based upon a ZIP+2 centroid where the street segment is matched and the ZIP+2 cluster centroid is the same for the entire street. The BG, tract and County are all the same for this street. Location is a ZIP+2 centroid.	yes
ZC5W	Location derived from a ZIP Code Centroid – unclassified Census accuracy – normally accurate to the county level – based on Post Office location (that delivers the mail to this address) where MORE then 80% of addresses in this ZIP Code are in a single census tract. REASONABLE Census ID accuracy. Location is assigned to the ZIP code centroid.	probably
Other Z		no

5.3. Geocoding of Contemporary (Visit 1-4) Addresses

Databases and programs used to generate geocodes vary and databases used by the same vendor are regularly updated. Therefore, we decided to resubmit Visit 1 and Visit 3 address with the Visit 2 and Visit 4 addresses so that all geocodes assigned would be obtained from a common source. The match rate for Washington County for Visit 1 was lower than that of other sites, as expected, due to the changes in addresses that occurred in the early 1990s. Table 3 presents data on the yield of Visit 1 to 4 addresses from our first submission to Vendor B.

Table 3: Results from Submission of V1-V4 ARIC Addresses to Vendor B

CODE from Vendor B	Visit 1	visit 2	visit 3	visit 4	Total
Nothing	19	17	1	13	50
A/AS	0	8	17	10	35
AS0	12826	13473	11665	10446	48410
AS1	174	161	138	131	604
AS2	42	48	48	44	182
AS3	2	0	0	0	2
To address: total	13044	13690	11868	10631	49233
To Block: total	75	295	323	379	1072
To Tract: total	117	92	48	41	298
ZC5W	227	32	20	18	297
ZC7N	1	4	1	0	6
ZC5X,5Y,5Z,9D,9G,9H (Not acceptable)	1982	495	253	228	2958
To County: total	2210	531	274	246	3261
Total	15465	14625	12514	11310	53914

Major reasons why addresses did not lead to a match included:
- Streets were renumbered between the time the participant provided the address and the time of geocoding
- Streets were renamed
- Route/street was moved or eliminated (e.g., US 29 is moved to a different route through town)
- Mailing address provided was not necessarily the place of residence (e.g., PO Box)

5.3.1. Efforts to Increase Visit 1 Match Rates in Washington Co.

Participants with Washington County Visit 1 addresses which did not geocode (n=998) were compared with the file originally described in Section 5.1. In cases where a geocode was assigned in the earlier geocoding effort (presumably due to supplemental manual geocoding, using maps and the help of an ARIC Washington County field center staff person), the geocodes assigned were used to replace the missing values in the more current data set. In the files containing the neighborhood SES variables, these participants were assigned the code 'A' for the variable TRTCODED.

During a review of these manually assigned addresses, we found that 60 Visit 1 Washington County participants with manually assigned tracts (trtcoded =A) had been placed in a tract in which the only building on record was a prison. Vendor B did not assign any of these participants' Visit 2 addresses to this tract; most were placed into one of two tracts adjoining the Visit 1 "prison" tract. We randomly selected 10 of these participants and asked the Project Coordinator in the Hagerstown ARIC field center to check reported addresses changes and moves using non-electronic files kept for this purpose. She determined that

although the addresses had "changed" the participants had not moved.

We tabulated the number of tract changes in Washington County between subsequent ARIC exams. Changes between Visits 1 and 2 were substantially greater (19%) than between Visits 2 and 3 (7%) and Visits 3 and 4 (7%). This difference and the questionable assignment of Visit 1 addresses to the "prison" tract raised some concern about the accuracy of the manual geocoding originally done. In order to investigate this problem, we identified 274 (7%) people in tracts where clusters of 5 or more appeared to move from a Visit 1 tract to an adjacent Visit 2 tract. We then enlisted the help of the Washington County ARIC field center staff who reviewed written records and found that 255 of the 274 people had not actually moved, although their address had changed. As a consequence of this review, we changed the Visit 1 tracts (which had been manually assigned) to the tract associated with the Visit 2 address. These people are identified in the data set with a trtcoded='c'.

5.3.2. Efforts to Increase Match Rates to Tracts by Data Cleaning

Other addresses in the data set that did not geocode were reviewed for obvious errors that might have resulted in a failure to geocode. Examples of common mistakes include spelling errors and a missing space between two words (MainST instead of Main ST). The obvious errors were corrected and these addresses were re-submitted to Vendor B.

5.4. Geocoding of Historical Addresses

Approximately 24,000 addresses representing places of residence at ages 30, 40 and 50 were edited and sent to Vendor B for geocoding.

Table 4: Geocoding Yield Of Final Submission Of Historical Addresses To Vendor B

Addresses expected		24225
Sent for Geocodeing		22176 (92%)
Returned with:	A code	16471 (74%)
	Usable Z code	145
	E or non usable Z code	5560
Addresses without usable geocoding:	address with street number	1693
	cross streets	925
	street name (or institution name)	2942

Major reasons for lack of geocodes included:
- Participant could not recall the complete address
- Streets were renamed or address numbering changed across time. (This problem was most common in Washington County. No efforts have been made to date to work with these addresses).
- Street or route moved or eliminated
- Zip codes not usable
- Street information misspelled or inaccurate (road instead of street, etc)

5.5. Quality Assessment of Geocoding

Assessment of the reliability and repeatability of commercial geocoding was done in four studies. Each of these investigations is reviewed below.

5.5.1. Comparing Geocodes Assigned by Two Vendors Five Years Apart

We compared the geocodes obtained for the 12,289 Visit 3 addresses sent to two different vendors (Vendor A, Vendor B) approximately five years apart. Of these, 500, or approximately four percent were not assigned the same geocodes by the two companies.

In a subset of 84 of these from Winston Salem we manually geocoded addresses to determine which geocode was correct and found that:
• 32 (38%) were in tracts assigned by Vendor B
• 5 (6%) were in tracts assigned by Vendor A
• 47 (56%) were on roads that divided the two tracts.

We followed up on the 47 IDs that were on roads that served as the border of tracts by using Yahoo files & TIGER/Line® files so that we could attempt to determine on which side of the road the addresses were located. On roads that fall on tract borders, side of the road is a determinant of tract assignment. According to an official of the US Bureau of the Census, the center of the road is considered the tract boundary; houses on one side of the street are assigned one tract and those on the other side are assigned an adjacent tract. Typically, this corresponds to even numbered addresses being assigned one tract while odd-numbered addresses are assigned the other tract. Our results are as follows:
• 36 (77%) fell in the tract assigned by Vendor B
• 3 (6%) fell in the tract assigned by Vendor A
• 7 (15%) could not be located

Our results on albeit a limited sample, favor the geocodes provided by Vendor B. However, given that the TIGER/Line® files used by the commercial geocoding companies are regularly updated it may be that the more favorable results seen for Vendor B are related to the files being more accurate at a later date and not the accuracy of the vendor per se.

5.5.2. Small Sample Short-term Repeatability of Vendor B

A set of 481 "complicated" Visit 3 addresses originally sent to Vendor B were resubmitted a few months later. The results are summarized in tables 5 & 5a below

Table 5: Comparison of longitudes & latitudes Assigned to 481 Addresses on two Submissions to Vendor B, Two Months Apart

Code	Frequency/%
Not assigned on either submission	5 (1%)
First submission only	110* (23%)
Same code on both submissions	357 (74%)
Different code on two submissions	9 (2%)

Table 5a: Statistics Pertaining to Nine Addresses Receiving Different Assignments on Two Submissions to Vendor B

Code	Located near	Located >.25 mile	Frequency
Exact address	2	6	8
Not exact address		Yes	1

* The resubmission took place after Vendor B had begun using 2000 files and an overlay method to assign 1990 tracts. These 110 addresses had originally been geocoded with z-code (zip + centroid) level accuracy which is not specific enough for the overlay method.

5.5.3. Large Sample Short-term Repeatability of Vendor B

We resubmitted a subset of 10362 historical addresses to Vendor B eight months after the original submission to the same geocoder. The results on repeatability across submissions are summarized in Table 6 below:

Table 6 Correspondence between FIPs Codes across 2 Submissions to Vendor B

Match state	Match county	Match tract	Frequency	Percent
no	No	no	1*	0.01
yes	No	no	3	0.03
yes	Yes	no	205	1.99
yes	Yes	yes	10110	97.97

* address was on the state line

5.5.4. Comparison of Geocodes Assigned by Vendors to a Gold Standard

A systematic assessment of repeatability and accuracy was conducted by our group (Whitsel EA, Rose KM, Wood JL, Henley AC, Liao D, Smith RL, Heiss G. Accuracy and repeatability of commercial geocoding in the LC-SES study. In

press, American Journal of Epidemiology). Briefly, the repeatability and accuracy of Vendor B was assessed. The 9-month repeatability of geocodes assigned by Vendor B to 1,032 participant addresses was uniformly high. Accuracy was assessed by comparing spatial coordinates associated with air pollution monitors that had spatial coordinates assigned by a gold standard method. These were located in the same areas as our study sites to those obtained by Vendor B. Match rates for addresses of EPA monitors were lower for Vendor B versus a second commercial geocoder (Vendor C) (76% vs. 88%). In contrast, discordance at the block group, tract and county level was greater for Vendor C vs.Vendor B. Coordinates assigned by Vendor C vs. B also were further from those in the EPA database . A published abstract can be found in *publications/presentations.*

6. Assignment of Historic Census Tracts

Given that census tract boundaries change across time, the 1990 census tract assignment provided by the geocoder could not be used to assign tracts to historic addresses. In order to abstract context-appropriate census information, we needed to find the correct tract for the appropriate year

6.1. Appropriate Census Year for Historic Addresses

We first determined the census year (1960, 1970, 1980) that corresponded most closely to the year in which a participant was age 30, 40 or 50 (Table 7).

Table 7: Assignment of Historical Census to Address by Birth Cohort and Age

Birth cohort	Age 30	Age 40	Age 50
'22-'25 N=1469	'52-'55 (1960 C**)	'62-'65 (1960 C)	'72-'75 (1970 C)
'26-'30 N=3349	'55-'60 (1960 C)	'65-'70 (1970 C)	'75-'80 (1980 C)
'31-'35 N=3708	'60-'65 (1960 C)	'70-'75 (1970 C)	'80-'85 (1980 C)
'36-'40 N=3978	'65-'70 (1970 C)	'75-'80 (1980 C)	'85-'90 (1990 C)
'41-'44 N=2230	'70-'74 (1970 C)	'80-'84 (1980 C)	'90-'94 (1990 C)

**Note: 1469 people turned 30 before 1956 and should have used 1950 census data: however, 1950 data is not available by tract so 1960 tract data is used.

6.2. Methods For Assigning Appropriate Census Year Tracts

We next determined the appropriate method of assigning historic addresses, depending on census year. There are two methods available:

Overlay method – The longitude and latitude points of an address are overlaid onto digitized files of tract lines (polygons). The advantage of this method is that any address that is assigned a latitude and longitude by the geocoder can be placed in a tract.

Comparability file method – The US Census Bureau maintains files which indicate how tract boundaries change from one census to the next. These are called comparability files. The advantage of this method is that accuracy is as good as the comparability files (we found only one error). The disadvantage is that when a more recent tract is made up of two or more tracts, or parts of two or more tracts from the previous decade, it is not possible to correctly determine the historical tract placement.

As a test, we compared the 1970 tract assigned by each method, using 13,044 addresses that were originally assigned latitude and longitude and 1990 tracts. Of these, 36% could not be assigned a 1970 tract using the comparability files because of identifiable tract splits between the periods. Of the remaining addresses (n=8348), 97% were assigned identical tracts by both methods.

With this test we felt confident using the overlay method when possible. Digitized census tract maps produced on CD ROM by Geolytics, Inc. were available for1970 and 1980. These map files, which include both map boundaries at the several geographic levels and census data from population and housing summary files, were available through Davis Library at the University of North Carolina at Chapel Hill. Latitudes and longitudes of the addresses that were successfully matched were linked with the appropriate census tract by historic census year.

These electronic polygon files were not available for 1960 so a combination of the overly and the comparability file method was used. Addresses from 1960 were assigned 1970 tracts using the overlay method then placed in the 1960 tracts using the file 'Tract Comparability: 1970 to 1960'.

6.3. Problems & Solutions with Assigning Historical Tracts

1. If the 1970 tract was made up of two or more 1960 tracts a person could not be placed in a 1960 tract without going back to actual addresses and attempting to place in a tract 'by hand'. This will be described in more detail below.

2. Jackson, MS & Washington Co. MD as well as much of the country were not tracked in 1960. For these areas we used the 1970 tracts.

3. There were 1,469 participants for whom the 1950 census would have been appropriate. As tract level census data was not available for 1950, the 1960 census was substituted, making gapyr as large as 8 years. See *Table 5* above. In one instance, a person was age 30 in 1952 and lived in a non-tracted area in 1960; therefore, the 1970 census was the only available source of data and the gapyr value is 18. (that is 18 years between yr lived that the address & the census used to describe the area of the address)

4. In order to use the overlay method described above, it was necessary to have longitude and latitude coordinates for an address obtained from an exact address

(match code starts with A). If the address was one of those with a match code starting with Z, coded to tract level, it was necessary to use the comparability file method to find the tract for the historic year.

5. When the historic addresses were ready for geocoding, Vendor B had already begun using the 2000 Census sources. We requested that they geocode to 1990 tracts and they agreed. Approximately 2 years after the coding was completed, we noticed an anomaly which led us to question the coding: for several streets which were identified tract boundaries, addresses with both odd and even dwelling numbers had been assigned to the same tract. Discussion with VENDOR B disclosed that they had used an overlay method to assign 1990 tracts to historic addresses previously coded to the 2000 census. We also discovered that, as is true for many commercial geocoders, they did not process the files in-house.

6. No electronic copy of 'Tract Comparability: 1970 to 1960' existed so we had to key in the print copy that is in the bound volumes of the '1970 census of Population and Housing' preceding each tracted area. These volumes were found Davis Library at the University of North Carolina at Chapel Hill, a Federal Depository Library

6.4. Manual Assignment to Tracts

There were several situations in which we attempted to hand tract participant addresses. This was most commonly done when the participant provided partial address information (street name but not number, cross streets, etc) and was also used for 1960 addresses when a 1970 tract mapped to two or more 1960 tracts. Detailed street maps of the four main study areas were obtained and census tract boundaries from the three historical censuses (1960, 1970 and 1980) were drawn by hand on these maps. Using standardized procedures we attempted to locate each address on the map. If the address could be located and was contained within the boundary of a tract, it was assigned the appropriate tract number. If for example an address was only a street that crossed multiple tracts or served as the boundary for two or more tracts, a census tract was not assigned.

A large number of Washington County, MD historical addresses were obsolete, because a major renumbering/naming of streets occurred in the early 1990's. We obtained detailed historical street maps from the Hagerstown Public Library and attempted to locate the original street names. We then attempted to hand geocode those found using the methods described above.

7. Census Based Socioeconomic Data

For 1970 and 1980, socioeconomic census tract data was obtained from Geolytics at the same time as the polygon tract files were obtained.

The 1960 census tract data was not available through Geolytics, and was obtained from electronic files available through the Odum Institute at UNC.

Much of the country however was not tracted in 1960 and has no electronic census data available. However, there are print files of socioeconomic housing data of towns and cities with a population greater then 10,000 by city block. This data was keyed and aggregated at the level of 1970 tracts boundaries for Jackson and Hagerstown.

For other non-tracted areas in 1960 we used the 1970 census data as the next best thing. A variable, 'gapyr', gives the gap between the year of the census data and the year the participant was a specified age. Optimally gapyr should have a value less than five. In those instances when 1970 census data must be substituted for 1960 data, the gap can be as great as 15 years.

8. Information on Data Sets

For a further description of these data sets, see *Data & Statistics* on this Website.

8.1. Individual SES Life Course Data Set

Lcdsc06a Date: 8/2004 – information from the ARIC visits & AFU
One record per person with variables selected from the following sources:
• Visit 1-4 data files: interview & exam every 3 years
• Annual Follow-up (AFU): phone call each year
• SES AFU – All the questions from the SES AFU are included except the questions about residence that were used to locate census tracts lived in at ages 30, 40, 50 and county at the age of 10 years

8.2. Neighborhood SES Data Sets
Lcdsc38a Date: 7/2004 – Historic neighborhood census data for ages 30, 40 and 50 was combined with Visit 1- 4 neighborhood data to create a dataset with neighborhood exposures for each decade from ages from 30-70.

Data set description: One record per person with census tract information for participant's address when they were 30-70 years old with up to 5 age points (age decades 30, 40, 50 ,60 70) over those years.   Censuses included range from 1960-2000.

Source of Address Data: All participants were asked to provide their address at age 30; addresses at ages 40 & 50 were selectively queried based on their age at the first ARIC visit. Among those who were already aged 40 or 50, respectively at Visit 1, we filled in tract information when available from the Visit 1-4 examinations by picking the Visit when the participant's age was closest to the missing age decade. If the participant had reached age 58 or 68 at the LC-SES interview, we also filled in tract information for ages 60 and 70, respectively.

Address data year gaps: The ideal would be to have address of residence at each age decade, as it falls on the calendar decade year.

Problem 1(addresses from visits 1-4 only): the date of the visits being used to fill in addresses did not necessarily fall on a participant's age decade, the visit with the age closest to the age decade being filled in was used & an age variable included with the actual age for the age decade be represented.

Problem 2 (all): the year the participant resided at the address did not usually fall on the census year so there is a variable (gapyr) created that is the number of years between the date of the address and the census year.

File construction & description of birth cohort: Because participants in the LC-SES study were born in different birth cohorts, the census most appropriate to a given age will not be uniform across the participants. Table 8 describes the distribution of census by age epoch.

Table 8: Census Used For Participants By Age Decade

Census	Age 30	Age 40	Age 50	Total
1960	7,123	1,122		8,245
1970	5,629	5,996	1,117	12,742
1980		1,903	1,390	3,293
Total	12,752	9,021	2,507	

The Neighborhood SES data has 23 variables (see *Table 9*) repeated for each of the 5 age categories. Because of the variation in source of characteristics by age, it was necessary to devise a meaningful convention for naming variables. The variable names of information for age 30, 40, 50, 60, 70 are prefaced with a3, a4, a5, a6 and a7 respectively, eg. a3unempl, a4unempl, and a7unempl are 'Percentage unemployed' at age 30, 40 & 70. Also each of the 5 points also has a variable CENSUS named a3census, a4census, etc. and gives the census year of the data used for that age. Adult age points with data taken from Visit 1 to 4 data also have a nonmissing age (a4age, a5age, .).

Quality of data: Optimally the gapyr is five or less, i.e., address is within 5 years of the decennial census data being used. However, to reduce the amount of missing data, we assigned census data from 1970 tracts to people in areas that in 1960 had no available census data. Because we were using 1960 census data (the first year with data compiled by tracts) for people turning 30 in the early 50s, it is possible to have a gapyr value as large as 18 years [e.g., one participant who turned 30 in 1952 and lived in an untracted area has 1970 data for age 30 (a3gapyr=18), age 40 (a4gapyr=8) and age 50 (a5gapyr=2)]. With use of the gapyr variables each investigator can selectively delete those participants for whom the interval from age at a particular address to when the census data was obtained is greater than what s/he feels is reasonably appropriate. Fortunately, the agegap is typically only a few years (i.e., the address is within a few years of the age being represented).

Table 9: Socioenvironmental Variables Available for Neighborhood SES by Census Year

	Childhood(County)			1960 non tract areas	1960* Tracted areas	1970**	1980***	1990****	2000#
	1930	1940	1950						
Income									
Median household income							X	X	X
Mean household income							X	X	X
Median family income			X				X	X	X
Mean family income				X	X	X	X	X	X
% Individuals below poverty						X	X	X	X
% Families below poverty			X						X
% of households with income interest, dividends, rent							X	X	X
% housing units that are owner occupied		X	X	X	X	X	X	X	X
Crops net $ amount	X	X	X						
Sales net $ amount	X	X	X						
Manufacturing net $ amount	X	X	X						
Crops, sales, Manuf. $ amount per person	X	X	X						
Education									
% adults 25+ yrs who have high school education		X	X		X	X	X	X	X
% adults 25 + with college degree		X	X		X	X	X	X	X
% 16-17 attending High School	X								
% 18-19 attending college	X								
% 10 & over illiterate	X								
Occupation									
% ages 16+ in professional, managerial, and executive		X	X		X	X	X	X	X
Percentage unemployed	X	X	X		X	X	X	X	X
housing									
% housing units occupied		X	X		X	X	X	X	X
% housing units dilapidated/deteriorating	X	X		X					
% occupied housing with > one person per room				X	X	X	X	X	X
% H units - Living in same house for last five years					X	X	X	X	X
Number of individuals living in household		X	X	X	X	X	X	X	X
Median value of owner occupied house	X	X	X					X	X
Mean value of owner occupied house				X	X	X	X	X	X
% households headed by a female *(with minor children)*						X	X	X	X
% households headed by a male *(with minor children)*							X		
% single parent homes *(men or women)*								X	X
% people in Urban area	X	X	X		x	X	X	X	X

* US Census tract level data, 1960
** censusCD70 - Geolytics
***censusCD80 – Geolytics
**** Census of population and housing, 1990   summary tape file 3 (STF3)
\# Census of population and housing, 2000   summary tract files (ICPSR)

8.3. Childhood Neighborhood SES Information

lcdsc53a Date: 11/2003 – There is one record per person with census information based on reported place of residence during childhood (linked to census closest to when the participant was age ten). All variables are at the county level, as tract level information was not available prior to 1960.

For the childhood data the variables are prefaced with a1.   Census years included range from 1930-1950.