APPENDIX AVAILABLE ON REQUEST

Special Report

Reanalysis of the Harvard Six Cities Study and the American Cancer Society Study of Particulate Air Pollution and Mortality

Part II: Sensitivity Analyses

Appendix A. Quality Assurance Audit of the Data

Daniel Krewski, Richard T. Burnett, Mark S. Goldberg, Kristin Hoover, Jack Siemiatycki, Michael Jerrett, Michal Abrahamowicz, Warren H. White, and Others

---

Correspondence may be addressed to Dr Daniel Krewski, Professor of Epidemiology and Statistics, Department of Epidemiology and Community Medicine, Room 3229C, 451 Smyth Road, University of Ottawa, Ottawa Ontario K1H 8M5, Canada

Although this document was produced with partial funding by the United States Environmental Protection Agency under Assistance Award R824835 to the Health Effects Institute, it has not been subjected to the Agency's peer and administrative review and therefore may not necessarily reflect the views of the Agency, and no official endorsement by it should be inferred. The contents of this document also have not been reviewed by private party institutions, including those that support the Health Effects Institute; therefore, it may not reflect the views or policies of these parties, and no endorsement by them should be inferred.

This document was reviewed by the HEI Health Review Committee but did not undergo the HEI scientific editing and production process.

UNIVERSITY OF OTTAWA

Faculty of Medicine **Faculty of Health Sciences**

Re-analysis of the Harvard Six-Cities Study and the American Cancer Society Study of Air Pollution and Mortality, Phase II: Sensitivity Analysis

Appendix A, B, C, D, E, F, G, H, and I

R. Samuel McLaughlin Center for Population Health Risk Assessment
Institute of Population Health
University of Ottawa

August, 2000

THE SIX CITIES STUDY

An independent team conducted a detailed audit of all data used in the analysis reported by the Original Investigators (Dockery et al 1993; referred to as the Part I data quality audit) as well as new variables used in our own sensitivity analyses. The Audit Team, headed by Ms Kristin Hoover, BS, MA, included Donna E. Foliart, MD, MPH, Warren H. White, PhD, and Ms Linda Calisti, BS. This team combined the resources and expertise of individuals who collectively have more than fifty years of experience in a wide variety of audits including university research, commercial analytic chemistry laboratories, toxicology, occupational exposure assessments and epidemiology (primarily in air pollution research). Members of this team have also been instrumental in the development of new audit methodologies for toxicological and epidemiological studies.

The Part I data quality audit was designed to provide an overview of the databases and an assessment of the data management procedures used by the Original Investigators. The audit also assessed the accuracy of data in the analytic files used in the original analyses relative to the original data from which they were derived. The objective of the Part II data quality audit was to evaluate the accuracy of the new variables selected by the Reanalysis Team for inclusion in their sensitivity analyses.

We have reported in detail the methods and results of the audit of the data used in the original analyses (see Part I report). In the present appendix, we provide a summary of the methods used in both phases as well as the results of the audit of the additional set of variables used in the sensitivity analyses. The sensitivity analyses used the original set of variables as well as those that have been collected and coded but not otherwise used in the published articles.

Methods

The data used in the Part I data quality audit were derived from 14 variables taken from questionnaires completed by the participants at their time of entry into the study, starting in 1974. We also conducted a detailed audit of the data that were used to estimate ambient air concentrations of particulate matter, sulfur dioxide, ozone and suspended sulfate that were measured in each community at one centrally located air monitoring station.

For both Parts I and II, we randomly selected 250 subjects whose questionnaires were used as the basis of the data quality audit. Part I also included another random sample of 250 death certificates that were used to audit the nosologic coding of the underlying cause of death and the date of death. The sample size of 250 was selected in order to provide reasonable statistical accuracy for achieving the goals of the data quality audit. Specifically, we selected this sample size to provide almost complete certainty of finding an error as small as 1% (Wang et al 1994), to distinguish between error rates of 1% and 5% with reasonable confidence, and to estimate error rates within about two percentage points of the true value. Further details are provided in Appendix A of the Phase 1 report.

For the Part II data quality audit, we included 17 variables from this initial questionnaire, five variables from follow-up questionnaires completed at three, six and 12 years after entry into the study (these were not used in the original paper) and two variables derived from measurements of pulmonary function conducted at time of entry into the study. In addition, for subjects selected for the questionnaire audit and who also died during the follow-up period, we also audited the underlying cause of death found on their death certificates.

The questionnaire, pulmonary function and death certificate data were audited against printouts of the electronic files. The audit for Part I consisted of comparing the information recorded in these sources to the data files used in the original analysis. For Part II, we made comparisons between original data and other data analysis files from the Original Investigators that were used subsequently in the sensitivity analyses.

Part I Audit

The data quality audit of the Six Cities Study encompassed more than 21,750 morbidity and mortality data points. Original questionnaires and death certificates were traceable via paper and electronic files, with the exception of one questionnaire (0.4%) and two death certificates (0.8%). All analysis files and supporting documentation for health and mortality data were available and traceable during the audit. The Original Investigators revised the form that served as the baseline questionnaire as the study progressed. Form 1-71 was used for Watertown, Harriman and St Louis. Form 77 (1–76) was used for Steubenville. For Topeka, Form 77 (1–76) was the baseline questionnaire used for some subjects, whereas Form 78 (1/77) was the first questionnaire used for other subjects. Form 78 (1/77) was used in Portage. Small discrepancies between these forms resulted in some study errors. Documentation of internal audits of the Six Cities Study from February and March of 1981 discussed these errors in detail and presented methods of correction (when feasible). Form 77(1–76) contained a misprint, so that for attained education the code "1" represented "grade school not completed". The older form 1–71 used code "1" to show that grade school was completed. Some interviewers crossed out the word "not" and coded this as "1" (representing "grade school completed") in order to make it consistent with the previous form. The original analysis was not affected by these changes, as only completion of high school was included in the statistical models.

The auditors discovered a computer programming error that resulted in the early censorship of time-on-study for approximately 1% of the reported person-years. The distribution of the loss of reported person-years was not equal in the six cities with the greatest accidental censorship in Portage and Topeka. The Reanalysis Team has carried out all their analyses using the appropriate dates of censoring and, as the Part I report showed, almost identical results were obtained. No change occurred in the person-years for Watertown.

One of the baseline questionnaires from the random sample of 250 could not be located, but follow-up questionnaires were available for this subject. Error rates in the analysis file from coding of the questionnaire administered at the time of entry into the study varied from 0% to 6% with most rates less than 2% with the exception of some occupational variables. Larger than average discrepancies were found

between the original data and the electronic database for the following variables: self-reported exposure to occupational dust (5.6%) and self-reported exposure to fumes and gases (6.0%).

The random sample of 250 death certificates selected for audit revealed that two death certificates were missing (0.8%). Of the remaining 248 available death certificates, eight discrepancies were noted among the five audited variables. There was a 100% correspondence between the original nosologist's coding and the ICD-9 code found in the analysis file. In two cases (0.8%) the auditor's attributed 4-digit ICD-9 code placed the death in a different analysis category than the code assigned by the study nosologist. Two errors were found when the date of death in the analysis file was compared to the dates recorded on the death certificates. One of these errors (year of death) was detected by the investigators following analysis and the current Six Cities Study database reflects the correct information. The second error (not corrected in the current file) involved the month of death. The analysis conducted by the Original Investigators and the Reanalysis Team utilized the month and year of death to define censoring times.

Mean or median levels of the pollutants were calculated for the period of interest for each of the six cities. Fine particle monitoring data were collected from 1979 to 1985, with the long-term average of all the daily fine particle data during this period used as an indicator of long-term exposure to particulate matter. The audit of the air pollution data revealed no problems with criteria pollutants and focused on the key explanatory variables used in the statistical analysis (concentrations of fine particle mass).

The dichotomous samplers used to collect fine and coarse particles were newly introduced instruments, and their field logs record a number of significant operational difficulties. Moreover, sample masses were determined in different years by two fundamentally different methods, and carried out by different organizations in different laboratories. Finally, the analyses of the dichotomous sampling data were not challenged with blind audit samples, as were the high-volume data.

Four distinct audit objectives for the dichotomous sampler data were established: 1) verify the reduction of primary measurements to concentrations; 2) evaluate procedures for validating and archiving measurements of concentrations of particles; 3) clarify the derivation of the published means; and 4) evaluate the sensitivity of these mean values to computational procedures and data selection criteria. Delays in locating records in the archives and involvement of multiple laboratories in the analyses of these data prevented the team from preselecting sets of dichotomous sampler data. As well, the data were not readily accessible and this placed a practical restriction on the data files that could be reviewed.

For the first objective, the auditor was able to verify the reduction of primary measurements to concentrations for the period November 1981 to January 1984, but was unable to achieve this objective for the other study years as the work was performed by the US EPA and records were not available in the archives of the Six Cities Study. Recalculated and reported values for fine and course mass concentrations were quite similar for the audited dataset (St Louis, May–July, 1983). While we were unable to directly validate data reduction from the EPA laboratory, we note that this laboratory was the leading practitioner of the methodology at that time.

A second objective of the audit was to reproduce the analysis dataset from the master database,

and this would provide verification of the criteria used to include or omit the data from the analysis. This objective could not be achieved because the original database no longer exists and no contemporary account of the criteria used to select data for analysis could be located. However, some criteria could be inferred by comparing the reconstructed analysis file with earlier records, and it was clear that different criteria were applied to different years. One example is rejection of observations with coarse/fine mass ratios outside a restricted range during the years 1979–1981 and inclusion of such observations in the period 1982 to 1985. This restriction does not bias the data in a predictable manner, and the empirical effect of the coarse/fine mass criterion on average concentrations was assessed by extending it for 1982 and later years, where it was not applied. The results of this exercise showed the effect on average concentrations occurred only in Topeka, where the average fine mass concentration calculated according to the criterion was 15% higher than that calculated from the uncensored data. Applying the criterion at Portage or Watertown produced no difference in average fine mass concentrations, even though it caused the rejection of over one third of the observations at these locations.

The final two objectives of the audit were to rederive the published long-term means (Dockery et al 1993) from the archival file of daily measurement data and evaluate their sensitivity to computational procedures and data selection criteria. The exploration of various averaging procedures and data selection criteria would determine the sensitivity of the analysis to mundane judgments of the investigators, and might clarify the specific procedures and criteria actually used in the original analysis. The value of this exercise was diminished, however, when in place of the original, comprehensive file of daily data it was necessary to substitute a reconstructed file pieced together from incomplete working copies derived through various procedures and selection criteria. Because the reconstruction was performed with knowledge of the published means, we cannot view a successful recovery of the published means from the reconstructed file as a fully independent validation. An alternative source of the dichotomous data were available in an analysis file used by Schwartz and collegues (1996) in their time-series study. This file contains, for the period 1979 to 1985, 22% more data points than the reconstructed file used in the original publication (Dockery et al 1993), and yields long-term average fine particle concentrations within 5% of the values in each city used in the original analysis.

Part II Audit

Variables for the Part II analysis were audited by conducting a comparison of selected variables from the baseline questionnaire, completed at the time of enrolment, as well as some other selected variables from the follow-up questionnaires to the data in the electronic analysis file provided to the Reanalysis Team. Underlying cause of death was evaluated using death certificates obtained by the Original Investigators for 60 subjects known to be deceased out of the 250 subjects selected to be in the random sample of audited questionnaires.

Variables obtained from the baseline questionnaire and audited in Part II were (SAS variable name from the analysis files in parenthesis):

- Date of Birth (DOB)
- rital Status (MARSTAT)
- Race (RACE)

- Occupation code from census (attributed using the 1980 Census scheme; OCC)
- Industry code (attributed using the 1980 Census scheme; IND)
- Number of years living in same town (YRSHERE1)
- Chest illness (bronchitis/emphysema/pneumonia) (CHSTIL1)
- Bronchial asthma (BRONAST)
- Alcohol consumption (DRINK)
- Beer consumption (BEER)
- Wine consumption (WINE)
- Liquor consumption (LIQUOR)
- Age started smoking (AGECIG)
- Number of packs of cigarettes smoked per week (CIGWK)
- Number of years of smoking cigarettes (YRSCIG)
- Underlying cause of death (COD) for any of the 250 individuals in the questionnaire random sample who are deceased and where death certificates were obtained (underlying cause of death for the 250 in the death certificate sample was audited for Part I; cause of death coded according to the Ninth Revision of the International Classification of Diseases (ICD-9))
- City of residence (CITY)
- Combination score for history of and/or treatment of heart trouble and high blood pressure (HBP)

Results Table A.1 shows a summary of the errors found in this second phase as well as some detailed comments regarding these errors. Details of the audit are provided below.

We found no errors for the variables date of birth (DOB), marital status (MARSTAT), city of residence (CITY) and race (RACE). Of the 250 questionnaires audited, one baseline questionnaire was missing, but data for these four variables were found on follow-up questionnaires.

No errors were found in the variable representing bronchial asthma (BRONAST) in the 249 subjects audited. The other audited chest illness variable (CHSTIL1) indicates whether the subject had bronchitis, emphysema, or pneumonia diagnosed by a physician. This variable was determined to have an error rate of 1.6% (four errors in 249 audited subjects). These errors were from two subjects with bronchitis and pneumonia being coded as having pneumonia only, one subject with bronchitis only was coded as having both bronchitis and pneumonia, and another subject with pneumonia was coded as "none". Our independent audit findings for this variable were lower than the internal audit conducted in 1981, in which three errors were found in 89 subjects sampled (3.4%). It was concluded in this internal audit that the error rate for this variable did not appear to have resulted from any systematic problem, and it was therefore decided that it was not necessary to conduct a recode.

When alcohol consumption from the baseline questionnaire was examined, no errors were noted in liquor (LIQUOR) and wine (WINE) consumption. For beer consumption (BEER), two errors were found out of 249 questionnaires (0.8%). One error was from one subject who was coded as having no consumption but should have been given a value indicating less than 200 ounces per week; another subject was coded with beer consumption of greater than 200 ounces per week, but should have been coded as having less than 200 ounces per week. For the present use of alcoholic beverages, one error (0.4%) was

observed in the coding of the variable DRINK. An error rate of 1.1% for beer consumption and 0% for present alcohol consumption was found in the 1981 internal audit.

The variables for cigarette consumption audited in Part II were: age started smoking (AGECIG), number of packs of cigarettes smoked per week (CIGWK), and number of years smoking cigarettes (YRSCIG). One error (0.4%) was found in the variable for age started smoking (AGECIG) in which the subject was found to have started smoking at age 15 versus age 40 in the analysis file. This appeared to be a simple shift in the figures during data entry, as the questionnaire clearly indicated that the subject smoked for 40 years, beginning at age 15. No errors were found when the Original Investigators audited this variable internally.

Records and internal audits at Harvard University showed inconsistencies in the calculation of cumulative smoking (as evaluated at date of entry into the study), arising from small discrepancies between the questionnaires used in the different cities. The method of calculation of cumulative lifetime of cigarettes (weekly amount times years of smoking) changed from Form 1-71 to Forms 77 and 78, thereby resulting in an underestimate of approximately 3% in Watertown, Harriman, and St Louis. There was a similar underestimate for this variable for former-smokers versus current-smokers in these same three cities. Our audit for the variable CIGWK (number of 20 cigarette packs smoked per week) showed that there were three errors in 249 audited subjects (1.2%). One subject smoked 140 cigarettes per week (7 packs per week, but was coded as smoking 14 packs per week). In three cases, 3.5 packs per week was rounded up to 4 and in another it was rounded back to 3 (the Audit Team counted this as one error because the rounding did not conform to the coding rules of the Six Cities Study). One subject reported smoking 10 cigarettes per week (equivalent to 0.5 packs per week) and this was reported as 3 packs per week. Our findings are consistent with Harvard University's internal audit that reported an error rate of 3.4% for this variable. Two errors (0.8%) were noted in the total number of years of cigarette smoking (YRSCIG). These were due to one subject who should have been attributed a duration of smoking of 31 years instead of the recorded 30 years. Another subject entry includes one year of abstinence that was coded as a year of smoking. Periods of time where the subject stopped smoking were to be coded and affected several smoking variables (LIFECIG), (YRSCIG), (CIGWK). Harvard University's internal audit of a smaller sample of eighty-nine revealed no errors.

The variable representing history of and/or treatment of heart trouble and high blood pressure was found on the baseline questionnaire (HBP). For Form 1-71, this information was contained in question 48; for form 77 (1–76), it was placed in question 60. The Audit Team found four errors in 249 subjects for an error rate of 1.6%. Harvard University's internal audit did not identify any errors.

Errors noted by the Audit Team included:

CASE	FORM	COMMENTS
1	1–71	Questionnaire indicated "enlarged heart," not treated; "yes" for treatment for hypertension. Auditor would code as 7 rather than 6
2	1–71	Narrative states "diabetic diet has corrected blood pressure

		problem, has taken no medicine". Auditor would code untreated high blood pressure, rather than "none" for high blood pressure
3	77 (1–76)	"Yes" circled for question regarding doctor informing patient of heart trouble, with "rheumatic fever" noted. Auditor would code untreated heart trouble rather than "none"
4	77 (1–76)	The questionnaire shows a diagnosis of high blood pressure, treated, with notation that hypertension was treated with "nerve pills". Auditor would code 6 for treated high blood pressure, rather than untreated high blood pressure

The Original Investigators used two variables to capture information on employment: industry of employment (IND) and occupation (OCC). For women coded as "housewife" in the industry variable, the husband's occupation was entered as the OCC variable. For industry code, five errors in 249 subjects (2.0%) were noted. Each entry is listed where the auditor questioned the file code as follows (auditor's code, followed by code in database):

CASE	COMMENTS ON INDUSTRY CODE (IND)
1	Auditor would code "717" (insurance). Questionnaire lists "ins co analyst" coded as "999" (unknown)
2	Auditor would code 0 (retired/unemployed) instead of code 888 (housewife) (auditor selected the "0" code as the woman's former job was recorded in the occupation variable)
3	Auditor would code 888 (housewife) instead of 0 (auditor selected the "888" code in this case as housewife was listed and the husband's job was recorded in the occupation variable)
4	Auditor would code 108 (sawmills, planning mills and millwork). Questionnaire lists lumber company, coded as 999 (unknown)
5	Auditor would code 338 (newspaper publishing and printing). Questionnaire lists newspaper, coded as 999 (unknown)

The findings of the Audit Team were consistent with the internal audit, in that the Original Investigators noted problems with coding housewives and those subjects who were retired or unemployed. Their error rate of 12.4% (11/89) in a smaller sample demonstrates that efforts were made to correct this variable (our error rate was 2%).

Audit of the occupation variable revealed five errors in 249 audited subjects (error rate of 2.0%). Each entry is listed where the auditor questioned the file entry:

CASE	COMMENTS ON OCCUPATION CODE (OCC)
1	Auditor would code 326 (insurance adjusters, examiners, investigators) Questionnaire lists "ins co analyst," coded 999, unknown)
2	Auditor would code 145 (teacher) Questionnaire lists school teacher, coded as 184

	(editors and reporters)
3	Auditor would code 903 (janitor) Questionnaire lists janitor, coded as 963 (marshals and constables)
4	Auditor would code 372 (secretary) Questionnaire lists case as currently unemployed, former job secretary, coded 282 (sales representative, wholesale, trade)
5	Auditor would code 622 (steelworker) Questionnaire lists spouse as steelworker, coded as 999 (unknown)

Harvard University's internal audit noted a number of discrepancies in the coding of occupations and industries (23.6% error rate) and it is clear from existing documents that considerable effort was expended in correcting these problems. In our Part I audit, error rates were highest in the self-reports of occupational exposures (5.6% error rate in the occupational dust variable and a 6.0% error rate for the occupational fumes and gases variable). These errors in exposures to dust and fumes are partly related to limitations of the earliest form of the questionnaire administered in Watertown, Harriman and St Louis.

The number of years resident in the city of enrolment (YRSHERE1) was coded by the Original Investigators from the baseline questionnaire. The Audit Team found an error rate of 2.0% (5/249), as follows:

CASE	Number of Years of Residency in City of Enrolment
1	File = 3, correct = 42
2	File = 6, correct = 24
3	File = 3, correct = 28
4	File = 40, correct = 72
5	File = 57, correct = 62

Three of these subjects had periods of time in other places and the earlier period of time in the city of interest was not included in the calculation. Instructions for coding (page 6, orange binder) were "All the time in the same town should be included, whether or not it is continuous." A similar type of error was noted in the internal audit conducted by the Original Investigators in 1981 (7.9% error rate for this variable). All of the errors noted in the sample from the internal audit were due to the total number of years in the same town not being counted, thus resulting in an underestimate. The total number of missing years represented by the five errors identified by our Audit Team was 119.

The underlying cause of death (COD) for the 250 individuals in the random sample of known deaths was audited in Part I. An additional audit was conducted of the 60 known deaths in the questionnaire sample of 250 subjects. Using information from the initial questionnaire, the auditor verified that the death certificate on file reflected the study participant. Personal identifiers (full name, Social Security number (SSN) (when available), birth date and gender) on the death certificate were audited against the information in the questionnaires. The audit verified that all of the death certificates pertained to the appropriate study participant. Using the International Classification of Diseases, revision-9 (ICD-9), Dr Foliart coded the underlying causes of death listed on the death certificate and compared them with the

nosologist's ICD-9 coding (noted on the pink cover sheet attached to the death certificate). No discrepancies were noted for any of the 60 deaths in this sample.

We audited pulmonary function variables from spirometry conducted at the time of enrollment into the study (forced vital capacity (FVC1) and forced expiratory volume (FEV11). During the audit, we found that the variables coded as FVC1 and FEV11 were derived from the participant's first "blow". The Original Investigators also generated a summary statistic for FVC1 and FEV11 that combined all attempts (varying from 5 to 8 blows); these new variables were referred to as optimum FVC (OPTFVC) and optimum FEV1 (OPTFEV1). The auditors focused on these latter values, as they are the most epidemiologically relevant. Due to time constraints, data from the analysis file for OPTFVC and OPTFEV1 for the 250 participants in the random sample, as well as the algorithm necessary to generate these values, were obtained directly from the Original Investigators. The Audit Team then checked the variables provided to the Reanalysis Team to determine that these values were identical to those provided by the Original Investigators while onsite.

The original algorithm used by the investigators to calculate OPTFEV and OPTFVC was not found. However, one member of the original study team provided a contemporary description of the most likely algorithm for calculation of these values. This entailed first identifying the three largest values. If the range between these values was less than 20 cubic centimeters (cc), the optimum value was defined as the mean of the three. If the range was 20cc or more, the mean of the two attempts within 20 cc was used. A correction factor for temperature was then applied to the mean. Thus, the final optimum values for FVC and FEV11 were the mean of three (or two) attempts, corrected for temperature.

Among the earliest participants in Watertown, a temperature correction factor was applied before the values for FVC and FEV11 were printed. For these participants, the Original Investigators removed the correction factor from each attempt, and then applied a correction factor to the mean of the three (or two) best attempts.

The Audit Team recalculated the mean FEV and FVC values using the above algorithm. As the Audit Team did not have access to the actual algorithm used in the study, all values were considered correct if they were between ± 5 cc of the reported values in the analysis file. No errors were noted in the 248 cases available in the audit sample. One subject was missing the baseline questionnaire and associated spirometry because the spirometry sheets were filed inside each baseline questionnaire. Another subject was missing spirometry data.

We audited five other variables not included in the Original Investigators' published paper, but which were derived from follow-up questionnaires completed at three and six years after the time of enrolment. These variables were:

- Height (HT) (in centimeters)
- Weight (WT) (in pounds)
- Smoking history (SMOK)
- Number of years of cigarette smoking (YRSCIG)
- Number of cigarettes smoked per week (CIGWK)

The audit of the analysis file for the height (HT) variable from the 3-year follow-up questionnaire revealed three errors in 249 questionnaires examined (1.2% error rate). Two subjects had entries for the third year switched with the sixth year and another subject had an incorrect entry for the year three questionnaire. Because two of the subjects in year three had values of height switched with year six, this caused an error in year six for these same two cases. The error rate for year six for height was 0.8%.

One rounding error was noted in year six when the weight (WT) variable was audited for the 3- and 6-year follow-up intervals. The error rate for year six was 0.4% (1/250). No errors were observed at the three year follow-up interval for any of the smoking variables (smoking status (SMOK), number of cigarette packs smoked per week (CIGWK), and number of years of cigarette smoking (YRSCIG)). No errors in smoking status (SMOK) were found at the 6-year follow-up. One rounding error was noted in year 6 for YRSCIG, thus resulting in an error rate of 0.4% (1/250). The number of cigarettes smoked per week (CIGWK) at the 6-year follow-up had one incorrect entry (0.4%; 1/250).

Also included in the Part II audit were three variables from the last follow-up questionnaire completed 12 years after study enrolment. These were:

- Height (HT) (in centimeters)
- Height (WT) (in pounds)
- Cigarettes per week (CIGWK)

A total of 247 questionnaires for year 12 were available, as three were missing from the participant's folder. No errors were observed in any of these variables with the possible exception of one case where height and weight appeared to be reversed on the questionnaire. The analysis file matched the questionnaire so this cannot be considered as an error in transcription but, rather, as an error in verification.

Of the 250 participants included in the audit sample, no errors were found in the identification of those individuals who moved or did not move outside the original city of residence, based on the residence histories coded by the Reanalysis Team. However, because some judgement was required in determining the date of this move (which had to be determined in some cases by the postal stamp on correspondence with the participants), there was some uncertainty as to the year of the first move outside of the original city of residence in 9 cases. Five of these nine discrepancies involved an error of only one calendar year. Although nine discrepancies were also noted in the month of the first move outside the original city of residence, these latter errors are expected to have a negligible effect on the analysis of population mobility.

Summary In this part of the audit, we found no errors that would induce important effects in the statistical analyses (under 5%), with the highest error rate being 2.4%. Although the error rate in the date of the first move outside the original city of residence based on residence histories coded by the Reanalysis Team was 3.6%, 5 of the 9 discrepancies involved an error of one calendar year in the date of the first move. We thus conclude that the data are of sufficiently high quality for the purposes of the Part II sensitivity analyses.

THE ACS STUDY

A similar data audit of the ACS Study was conducted using data from the reduced CPS-II cohort, as described by Pope and collegues (1995). There were three main differences, however, between the audit of this study and the Six Cities Study. First, the SAS data files used in the original analysis were not available. It was thus necessary for the ACS to reconstruct these datasets to correspond to the analytic files used by the Original Investigators. Second, personnel who were involved in the original formulation and conduct of the ACS CPS-II were no longer available to answer detailed questions regarding the procedures for data collection and data management. Third, significant amounts of documentation for the ACS Study were lost when the ACS moved their main office in New York City to Atlanta. Thus, in comparison to the Six Cities Study, there was less documentation available to audit each variable; the auditable information and data were limited to microfilmed death certificates, microfilmed questionnaires, and some computer programming information. As was reported in Part I, documentation of the ascertainment of vital status during the follow-up no longer exists nor does detailed information on the coding of each variable. Thus, the coding rules were often determined by the Audit Team through inference instead of documentation.

As with the audit of the Six Cities Study, we randomly selected 250 questionnaires and 250 death certificates for audit. Microfilm copies of questionnaires and death certificates were traceable with the exception of one questionnaire (0.4%) and eight death certificates (3.2%). Two additional death certificates were traced but the causes of death were not legible.

Part I Audit

During the course of the reconstruction of the analysis files, it was found that deaths in women occurring between September 1, 1988 and December 31, 1989 (end of follow-up) had not been included in the original analysis. These data were given to the Reanalysis Team enabling us to examine the impact of including the additional 5,421 female deaths that occurred during this period. The Original Investigators reported that there was no change in the overall study results when these additional data were added. In addition, it was found that due to a mistake in computer programming, deaths attributed to asthma were not included among the rubric of cardiopulmonary deaths. This did not affect the results of the study because the number of asthma deaths was relatively small.

The data quality audit demonstrated that the values for the variables included in the electronic data file used by the Original Investigators were in good agreement with the values on the original questionnaires. The review of variables drawn from the questionnaire sample included study identification number, race, sex, age, smoking history (eight variables), passive smoke exposure (three variables), alcohol (three variables), selected occupational exposures (six variables), education, height, weight, time-on-study, vital status, and month and year of death (when applicable).

The records of the determination of vital status as conducted by the ACS volunteers were lost when the ACS relocated to Atlanta. Time-on-study was therefore recalculated by the auditors assuming that those individuals indicated as being alive (through the vital status variable) were alive until the end of the study. Vital status of the 250 subjects in the sample of questionnaires was audited against three sources: a search of the National Death Index (NDI) from 1982–1989, a review of participants in an independent

nutritional survey conducted by the ACS after 1989, and a search of the Social Security database (available through the Internet). No discrepancies in vital status were found.

The review of the random sample of 250 death certificates found several inconsistencies. One death certificate did not pertain to the study participant (0.4%) and two errors in date of death were found (0.8%). In 15 of the 240 death certificates with legible causes of death (6.3%), the auditor's two- or four-digit code did not match the code in the analysis file. In four cases (1.6%), the auditor's four-digit ICD-9 code would place the death in a different analysis category as compared to the code assigned by the study nosologist.

During the review of the death certificates, an additional computer programming error was detected: the statistical program used to group causes of death placed two codes of cardiovascular deaths into the "other deaths" category. The ACS staff was notified of this programming error and a review of the complete cohort of deaths was performed. The two codes accounted for only 71 deaths among the total cohort, and the reassignment of these deaths to the cardiovascular category did not affect the final results (as reported by the ACS to the Audit Team).

The audit of the air pollution data were significantly more problematic than other study variables. No new air pollution data were gathered specifically for the ACS Study to respond to the objectives of the air pollution analyses nor were the original questionnaires designed specifically for the purpose of examining the association between long-term exposure to particulate air pollution and mortality. Rather, Pope and collegues (1995) designed this study to take advantage of existing databases; viz. the ACS cohort and concentrations of air pollutants that had been widely circulated and used previously by other investigators. The air pollution data are incompletely documented and are based on data that are now technologically difficult to access. Thus, in the absence of access to the original sources of the air pollution data, it was not possible to audit the data tracing through each of the steps to the resultant final dataset used in the statistical analyses (ie, instrument operating logs, measurement of filter weights, inclusion/exclusion criteria, and data processing).

The original derivation of annual average ambient sulfate levels was not documented and could not be audited. Instead, the averages used by the Original Investigators were compared with new values independently derived from daily data retrieved from official EPA data records. The results of this activity are discussed in another section of this Part II report. The fine particle levels were confirmed to have been correctly extracted from the original source, a technical report published by Brookhaven National Laboratory (Lipfert et al 1988).

Part II Audit

Variables for the Part II analysis were audited by conducting a comparison of the data from baseline questionnaires, completed at the time of enrolment, to data in the electronic analysis file provided to the Reanalysis Team. Variables (in alphabetical order) obtained from the baseline questionnaire and audited in Part II were (SAS variable name from the analysis files in parenthesis):

- Arthritis (ARTHRTIS)

- Asbestos exposure (in years) (ASBEYRS)
- Asthma (ASTHMA)
- Bladder disease history (BD)
- Beer consumption (previous) (BEERPR)
- Beer consumption (years of previous consumption) (BEERPRYR)
- Chronic indigestion (CI)
- Cirrhosis of liver (CL)
- Coal/stone dust exposure (in years) (COALDYRS)
- Coal tar pitch asphalt exposure (in years) (COALTYRS)
- Colds/flu (in past year) (COLDS)
- Colon polyps (CP)
- Cysts of breast (women only) (CYSTS)
- Diabetes (DBT)
- Diverticulosis (DC)
- Diesel engine exhaust (years of exposure) (DIESYRS)
- Duodenal ulcer (DU)
- Emphysema (EMPHYS)
- Ever smoked cigarettes at least one per day for one year's time (EVERSMK)
- Exercise (amount) (EXERCISE)
- Formaldehyde exposure (in years) (FORH_YRS)
- Gall stones (GST)
- Gynecologic problems diagnosed by physician (women only) (GYN)
- High blood pressure (HBP)
- Heart disease (HD)
- Hepatitis (HEPTS)
- Hay Fever (HF)
- Heart medicine (years of consumption) (HRTB)
- Heart medicine (monthly consumption) (HRTX)
- Kidney disease (KD)
- Kidney stones (KS)
- Liquor (previous) (LIQPR)
- Liquor (previous consumption in years) (LIQPRYR)
- Last occupation/retired (L_OCCUP)
- Marital status (MARITAL)
- Occupation (current) (OCCUP)
- Occupation (total years in current occupation) (OCCUPYR)
- Occupation (longest occupation) (OTH_JOB)
- Occupation (total years for longest occupation) (OTH_YRS)
- Other medical conditions (OTHER)
- Prostate problems (men only) (PROSTR)
- Rectal polyps (RP)
- Stroke (ST)
- Stomach ulcer (SU)

- Tuberculosis (TB)
- Thyroid medication (in years) (THYRB)
- Thyroid condition (THYROID)
- Thyroid medication (monthly consumption) (THYRX)
- Tylenol™ (in years) (TYLENOL)
- Tylenol™ (monthly consumption) (TYLENOL)
- Water (source of drinking water) (WATER)
- Water additives in drinking water (WATERADD)
- Wine (previous consumption) (WINEPR)
- Wine (years of previous wine consumption) (WINEPRYR)
- Years of residence in present neighborhood (YRSLIVE)

Results Table A.2 shows a summary of the errors found in this second phase as well as some detailed comments regarding these errors. Details of the results of the audit are provided below.Microfilm copies of questionnaires were traceable with the exception of three questionnaires (1.2%) out of the 250 included in the randomly selected sample for audit. As records were lost when the ACS moved their head office to Atlanta, the Audit Team had to determine most of the coding conventions by inference as opposed to the Six Cities Study where documentation could be examined. In general, the Audit Team found an inconsistent use of blanks, zero values, and "dots" throughout. In the instances in which the Audit Team was able to determine that these inconsistencies were not likely to affect the analysis file, these discrepancies were not counted as errors. Out of the 55 variables audited for the sensitivity analyses, no errors were observed in 34 of these variables. These variables where no errors were found included: arthritis, asbestos, bladder disease, beer consumption (previous amount and years), chronic indigestion, cirrhosis of the liver, coal/stone dust exposure, coal tar pitch asphalt exposure, colon polyps, breast cysts, diabetes, diverticulosis, diesel engine exhaust, duodenal ulcer, emphysema, exercise, formaldehyde exposure, gall stones, gynecologic problems, heart disease, heart medicine (month and yearly), prostate problems, rectal polyps, stroke, stomach ulcer, tuberculosis, thyroid medication (yearly), Tylenol™ (monthly and yearly), water additives in drinking water, wine (previous years) and years resident in present neighborhood.

Nine variables had one error (0.4%) and these were: asthma, colds, high blood pressure, hepatitis, previous liquor consumption in years, marital status, thyroid (both condition and medication variables) and previous amount of wine consumption. Two errors (0.8%) were found in each of the following five variables: hay fever, kidney disease, previous amount of liquor consumption, total years in current occupation and the variable for other medical conditions not otherwise coded. Variables for kidney stones and drinking water source had three errors each (1.2%). The variable for subjects who ever smoked at least one cigarette per day for one year's time, (EVERSMK) had an error rate of 2.0% based on five subjects (of 247) who should have been coded as smokers versus the blanks found in the analysis file. Eight errors (of 247) were found in the coding of total years for longest occupation held by each subject (OTH_YRS). In four cases, errors resulted from years in the longest occupation being coded as a dot in the file instead of the actual value.

When the variable representing last occupation/retired (L_OCCUP) was audited, two different categories of errors were found. For 103 cases (42%), the questionnaire was blank for the variable

(L_OCCUP), but the questionnaire did contain an entry for the variable "OTH_JOB" (occupation held for the longest period of time). The coder entered the code that was used for "OTH_JOB" in both columns, "L_OCCUP" and "OTH_JOB". The Audit Team consulted with ACS to determine if this procedure was part of the coding instructions, but they were unable to verify the original instructions. When the variable for "last occupation/retired" was audited after consideration of the previously described coding issue, 18 errors were noted in 247 questionnaires (7.3%). The most common error was the use of the code "Other" when a more specific code was appropriate. Other errors included two mechanical engineers who were coded as mechanics, a cafeteria worker was coded as an office worker, a bank teller was coded as an office worker, a receptionist was coded as an auto mechanic, a printer was coded as a clergy, and a teacher was coded as a nurse. One subject who worked previously for an artistic carton company was coded as an architect, and a former postal worker was not assigned a code.

We found that the variable for current occupation (OCCUP) had a total of 39 errors in 247 questionnaires (15.8%). In 20 of these 39 cases, occupation was coded as "Other" when existing occupational categories allowed for more specific coding of the entry. Eight errors involved failure to properly code retired individuals. Other errors involved coding a sewing machine operator as a machinist instead of a sewer; two other cases were coded as clergy instead of as school principal; one housewife was coded as a telephone operator; one subject was coded as employed in banking when the reported occupation was a freelance writer; one childcare worker and one laborer were coded as technician (radio, x-ray, dental or laboratory); one housewife was coded as a steel mill worker; one homemaker and one part-time office worker were coded as working in the legal profession; and one psychologist should have been coded as a therapist. There was one instance where a licensed practical nurse (code 33) was coded as a registered nurse (code 24).

The Original Investigators also coded the occupation held for the longest period of time (OTH_JOB). We found 20 errors in 247 questionnaires (8.1%). In four cases, the code for "Other" was used when a more specific code was appropriate. In three cases, the questionnaire entry was blank, but a code was entered into the database. In two cases, the questionnaire had an entry, but the database was blank. Other errors involved a dry cleaner coded as a bus driver, a painter coded as clergy, two real estate agents were coded as working in sales or coded as teacher (two cases), an engineer was coded as a teacher, an insurance agent was coded as a manager, a clerical worker was coded as a steel mill worker, a laborer was coded as a lab technician, a dairy plant worker was coded as farmer, a receptionist was coded as a carpenter, and an office worker was coded as a welder.

Summary In this part of the audit, for the non-occupational variables, we found no errors that would induce important effects in the statistical analyses (under 5%), with the highest error rate being 3.2%. We found very large discrepancies, however, in the coding of occupation and industry, with error rates for last occupation/retired of 7.3%, current occupation of 15.8%, occupation of longest employment of 8.1%, and total years of employment in longest occupation of 3.2%. We thus conclude that the non-occupational data are of sufficiently high quality for the purposes of the Part II sensitivity analyses but that the use of the occupational data may induce unacceptable errors in the analyses.

REFERENCES

Dockery DW, Pope CA, Xu X, Spengler JD, Ware JH, Fay ME, Ferris BG, Speizer FE. 1993. An association between air pollution and mortality in six US cities. New England J of Medicine 329 (24):1753–1759.

Lipfert FW, Malone RG, Daum ML. 1988. A statistical Study of the Macroepidemiology of Air Pollution and Total Mortality. Brookhaven National Laboratory, April, BNL 52122 US-404.

Pope CA, Thun MJ, Namboodiri MM, Dockery DW, Evans JS, Speizer FE, Heath CW. 1995. Particulate air pollution as a predictor of mortality in a prospective study of US adults. Am J Respir & Crit Care Med 151:669–674.

Schwartz J, Dockery DW, Neas LM. 1996. Is daily mortality associated specifically with fine particles? Journal of the Air & Waste Management Association. 46:927–939.

Wang Y, Krewski D Bartlett S, Zielinski JM. 1994. The Effect of Record Linkage Errors on Statistical Inference in Cohort Mortality Studies. Technical Report 245. Laboratory for Research in Statistics and Probability, Carleton University, Ottawa, Ontario, Canada.

Table A.1. Harvard Six-cities Study: Findings from the Phase II Audit of the Original Study Questionnaire[1].

Variable (SAS variable name from the analysis files)	Description of Variable	No. of Errors Found/No. of Questionnaires Examined	% Errors	Error rate from Harvard University's Internal Audit (1981)	Comments from the Phase II audit
AGECIG	Age started smoking; 0 = non-ers; ages 1-5 allowed by coding	1/249	0.4	0 /89 (0.0%)	Started smoking at age 15 not age 40 as coded
BEER	Beer (oz./wk.); 0 = none, 1= < 200 oz/wk, 2 = > 200 oz/wk	2/249	0.8	1/89 (1.1%)	One subject coded as 0 (none) but should have been 1 for < 200 oz/wk. One subject coded as 2 (> 200 oz/wk) but should have been coded 1 for < 200 oz
BRONAST	Bronchial asthma; 0=no, 1= yes at present, 2= yes in past, but not now	0/249	0.0	0/89 (0.0%)	
CHSTIL1	Chest illness diagnosed by doctor; 0=no for bronchitis, emphysema or pneumonia, 1= yes for bronchitis, 2=yes for emphysema, 4 = yes for pneumonia;	4/249	1.6	3/89 (3.4%) Harvard's audit concluded that error rate for this variable did not appear to have resulted from any systematic problem so no	Two subjects with bronchitis and pneumonia coded as pneumonia only; one subject with bronchitis only was coded as having both bronchitis and pneumonia; one subject had

Table A.1. Harvard Six-cities Study: Findings from the Phase II Audit of the Original Study Questionnaire[1].

	higher numbers from subjects diagnosed with two or more diseases			re-coding was done	pneumonia that was coded as none
CIGWK	Number of packs of cigarettes smoked per week (20 cigs/pk)	3/249	1.2	3/89 (3.4%)	One subject smoked 140 cigarettes/wk which is 7 pks/wk, but was coded as smoking 14 pks/wk, in three cases, 3.5 pks/wk was rounded up to 4 and in another it was rounded back to 3, one subject reported 10 cigarettes/wk which would be 0.5 pk/wk and this was reported as "3" pks/wk
CITY	City of Enrolment	0/250	0.0	0/89 (0.0%)	
DOB	Date of birth	0/250	0.0	0/89 (0.0%)	Differences in month or year of birth were noted in two instances, but established coding rules allowed data to be used from a later questionnaire

Table A.1. Harvard Six-cities Study: Findings from the Phase II Audit of the Original Study Questionnaire[1].

					that matched database entries.
DRINK	Present use of alcoholic beverages; 0=no, 1=yes and part B asks if use is as often as 1 day/wk; 0=no, 1=yes, 2= sum of both yes scores	1/249	0.4	0/89 (0.0%)	One subject was entered as 1 and should have been coded as 2
HBP	Heart/Blood Pressure Trouble (Subjects asked if doctor ever diagnosed high blood pressure or heart problems. If yes, had this been treated in the last ten years. Scores could total a high of "8")	4/249	1.6	0/89 (0.0%)	One entry not coded for treatment medication, one entry should have been coded for high blood pressure, two subject entries should have included the previously diagnosed heart problems.
IND	Industry code	5/249	2.0	11/89 (12.4%) Harvard's internal audit stated that retired/disable dunemployed could not be distinguished	Each entry is listed where the auditor questioned the file entry: Case 1: 717 vs. 999. Questionnaire lists "ins co

Table A.1. Harvard Six-cities Study: Findings from the Phase II Audit of the Original Study Questionnaire[1].

				and many errors in interpretation in this variable existed. Two common errors: – One error is that working wives were often coded as housewives without reference to outside employment – Second Error unjustified assumptions are made about jobs when no information is available as to specific duties. Harvard's documentation shows that work was conducted to correct several of these coding problems.	analyst" Case 2: 0 (retired/unemployed) vs. 888 (housewife) (auditor selected the "0" code as the woman's former job was recorded in occupation variable) Case 3: 888 vs. 0 (auditor selected the "888" code in this case as H/W was listed and the husband's job was recorded in the occupation variable.) Case 4: 108 vs. 999. Questionnaire lists lumber company. Case 5: 338 vs. 999. Questionnaire lists newspaper.
LIQUOR	Liquor (oz./wk.); 0 = none, 1= < 1-25 oz/wk, 2 = 26+ oz/wk	0/249	0.0	2/89 (2.2%)	

Table A.1. Harvard Six-cities Study: Findings from the Phase II Audit of the Original Study Questionnaire[1].

MARSTAT	Marital status	0/250	0.0	1/89 (1.1%)	Marital status could be assessed from a follow-up questionnaire
OCC	Occupation code; documents show that this variable was later superceded by COROCC	5/249	2	When Harvard audited the COROCC code, 21/89 errors noted (23.6%). Documents show efforts to correct errors.	Each entry is listed where the auditor questioned the file entry: Case 1: 326 vs. 999 (unknown) Questionnaire lists "ins co analyst" Case 2: 145 vs. 184. Questionnaire lists "school teacher" Case 3: 903 vs. 963. Questionnaire lists "janitor" Case 4: 372 vs. 282. Questionnaire lists case as currently unemployed, former job "secretary" Case 5: 622 vs. 999. Spouse is listed as steelworker Case 6: ORNL (? Code) vs. 561 (code not found)

Table A.1. Harvard Six-cities Study: Findings from the Phase II Audit of the Original Study Questionnaire[1].

RACE	Race; Codes existed for other races, but all study participants were Caucasian (code=1)	0/250	0.0	0/89 (0.0%)	
WINE	Wine consumption (oz/wk); 0=none, 1=< 99, 2=100+	0/249	0.0	0/89 (0.0%)	
YRSCIG	Total years smoked cigarettes	2/249	0.8	0/89 (0.0%)	One subject should have been coded "31" instead of "30." One entry includes one year of abstinence.
YRSHERE1	Number of years resident in this town	5/249	2.0	7/89 (7.9%) Internal audit noted that a consistent set of coding rules had not been carefully followed and that years in service should have been subtracted. Years in same city is counted even if not continuous	Errors noted were: Case 1 (file=3, correct=42), Case 2 (file=6, correct=24), Case 3 (file=3, correct=28), Case 4 (file=40, correct=72), Case 5 (file=57, correct=62)

[1] Note that variables for date of birth (DOB), marital status (MARSTAT), race (RACE) and city (CITY) were taken from Form 85 (9/87) because one subject file was missing the initial questionnaire for the study.

Table A.2. The American Cancer Society Study: Findings from the Phase II Audit of the Original Study Questionnaire[1]

Variable (SAS variable name from the analysis files)	Description of Variable	No. of Errors Found/No. of Questionnaires Examined	% Errors	Comments
ARTHRTIS	Arthritis diagnosed by physician	0/247	0	
ASBEYRS	Asbestos exposure (in years)	0/247	0	
ASTHMA	Asthma diagnosed by physician	1/247	0.4	The questionnaire indicated the subject had asthma, but it was coded as absent in the analysis file.
BD	Bladder disease diagnosed by physician	0/247	00	
BEERPR	Beer consumption (previous)	0/247	0	
BEERPRYR	Beer consumption (years of previous consumption)	0/247	0	
CI	Chronic indigestion diagnosed by physician	0/247	0	
CL	Cirrhosis of liver diagnosed by physician	0/247	0	
COALDYRS	Coal/stone dust exposure (in years)	0/247	0	
COALTYRS	Coal tar pitch asphalt exposure (in years)	0/247	0	

Variable (SAS variable name from the analysis files)	Description of Variable	No. of Errors Found/No. of Questionnaires Examined	% Errors	Comments
COLDS	Colds/flu (number of times subject had colds or flu in the past year)	1/247	0.4	One cold was reported by the subject, but the analysis file said "0."
CP	Colon polyps diagnosed by physician	0/247	0	
CYSTS	Breast cysts diagnosed by physician (women only)	0/247	0	
DBT	Diabetes	0/247	0	
DC	Diverticulosis diagnosed by physician	0/247	0	
DIESYRS	Diesel engine exhaust (years of exposure)	0/247	0	
DU	Duodenal ulcer diagnosed by physician	0/247	0	
EMPHYS	Emphysema	0/247	0	
EVERSMK	Ever smoked cigarettes at least one per day for one year's time (note that the male questionnaire includes cigars and pipes in this question)	5/247	2.0	Five subjects should have been coded as smokers versus the blanks found in the analysis file.
EXERCISE	Exercise (amount of exercise	0/247	0	

Variable (SAS variable name from the analysis files)	Description of Variable	No. of Errors Found/No. of Questionnaires Examined	% Errors	Comments
	through work or play characterized as none, slight, moderate or heavy)			
FORH_YRS	Formaldehyde exposure (in years)	0/247	0	
GST	Gall stones	0/247	0	
GYN	Gynecological problems diagnosed by physician (women only)	0/247	0	
HBP	High blood pressure diagnosed by physician	1/247	0.4	One case was coded as having high blood pressure when the questionnaire indicated the person did not have this condition.
HD	Heart disease diagnosed by physician	0/247	0	
HEPTS	Hepatitis	1/247	0.4	One case should have been coded as having hepatitis, based on questionnaire entry.
HF	Hay Fever diagnosed by physician	2/247	0.8	The questionnaire indicated that two subjects had hay fever, but this condition was absent in the analysis file for both cases.
HRTB	Heart medicine (years of consumption)	0/247	0	
HRTX	Heart medicine (monthly consumption)	0/247	0	
KD	Kidney	2/247	0.8	One case was lacking the code

Variable (SAS variable name from the analysis files)	Description of Variable	No. of Errors Found/No. of Questionnaires Examined	% Errors	Comments
	disease			for kidney disease in the analysis file and the other case was incorrectly coded as having this disease.
KS	Kidney stones	3/247	1.2	Two cases should be coded as having kidney stones, based on questionnaire entry. Conversely, another case should not have been coded as having this condition
LIQPR	Liquor (amount consumed previously)	2/247	0.8	Two drinks per week on questionnaire was entered as 97 in one case and a dot in another in the analysis file.
LIQPRYR	Liquor (previous consumption in years)	1/247	0.4	Fifteen years of consumption was reported on the questionnaire, but a dot appears in the analysis file.
L_OCCUP	Last occupation/retired	18/247 103/247 (analysis file contained entries that matched an adjacent, related column when the questionnaire itself was blank, thus, pertinent information existed on the questionnaire, but it was not recorded in the proper place.)	7.3 42	Examples of miscoded last occupations were: mechanical engineer coded as a mechanic (2 cases), cafeteria worker coded as an office worker, bank teller coded as an office worker, receptionist coded as an auto mechanic, printer coded as clergy and a teacher was coded as a nurse. One subject who previously worked for an artistic carton company was coded as an architect and a former postal worker was not assigned a code. For one case, the database had an entry, but the questionnaire was blank for variables L_OCCUP and OTH_JOB. For another, the questionnaire was blank and the subject was coded as

Variable (SAS variable name from the analysis files)	Description of Variable	No. of Errors Found/No. of Questionnaires Examined	% Errors	Comments
				unemployed. In 7 cases, the code for "other" was used when more specific codes existed. Questionnaires in these 7 cases listed employment as a guard, a secretary, an orthopedic surgeon, laborer, accountant, bus driver, and government worker.
MARITAL	Marital status	1/247	0.4	One subject should have been coded as single, but was coded as married.
OCCUP	Occupation (current)	39/247	15.8	In 20 of 39 cases, occupation was coded as "Other" when coding rules allowed more specific occupations to be coded. Examples of questionnaire entries coded as "Other" include warehouseman, school superintendent, real estate agent, librarian, housewife, legal arbitrator school administrator, print shop worker, disabled, security guard, executive (2 cases), lithography, machine operator, professor (2 cases), bus driver, accountant, seaman, and educator. Eight cases involved failure to properly code retired subjects. The remaining 11 cases involved: Sewing machine operator coded as a machinist instead of a sewer. School principal coded as clergy (2 cases). Housewife coded as a telephone operator. Freelance writer coded as in banking profession. Child care worker

Variable (SAS variable name from the analysis files)	Description of Variable	No. of Errors Found/No. of Questionnaires Examined	% Errors	Comments
				and a laborer coded as technician (radio, x-ray, dental or laboratory). Housewife coded as a steel mill worker. Homemaker and part-time office worker coded in the legal profession. Psychologist should have been coded as a therapist. Licensed practical nurse (code 33) coded as registered nurse (code 24).
OCCUPYR	Occupation (total years in current occupation)	2/247	0.8	One case should have been coded as 35 years instead of 32. Another case should have been coded as 40 years versus a dot on the form.
OTH_JOB	Occupation (longest occupation)	20/247	8.1	In four cases, the occupation was coded as "Other" when a more specific code applied. Examples of questionnaire entries coded as "Other" included rural school teacher, police sergeant, lithographer, and accountant (CPA). In four cases, the questionnaire was blank, but the database had an entry. In two cases, the questionnaire had an entry, yet the database was blank. The remaining 11 cases involve: dry cleaner coded as bus driver, painter coded as clergy, real estate agent coded as sales or teacher (2 cases), engineer coded as teacher, insurance agent coded as manager, clerical worker coded as steel mill worker, laborer coded as lab technician, diary plant worker

Variable (SAS variable name from the analysis files)	Description of Variable	No. of Errors Found/No. of Questionnaires Examined	% Errors	Comments
				coded as a farmer, receptionist coded as a carpenter, and office worker coded as a welder.
OTH_YRS	Occupation (total years for longest occupation)	8/247	3.2	One case had no entry on the questionnaire, but the analysis file had eleven years which the amount of time the person had been retired. In four cases, a dot appears in the analysis file when the questionnaire listed a specific number. Six and one half years was incorrectly rounded to 8 years. Ten years of nursing duties should have been recorded versus the five in the analysis file. One subject should have been coded as having 45 years versus the 12 years found in the analysis file.
OTHER	Other medical conditions	2/247	0.8	One illness was not coded and another had the wrong code.
PROSTR	Prostate problems (men only)	0/247	0	
RP	Rectal polyps diagnosed by physician	0/247	0	
ST	Stroke	0/247	0	
SU	Stomach ulcer diagnosed by physician	0/247	0	
TB	Tuberculosis	0/247	0	
THYRB	Thyroid medication (in years)	0/247	0	
THYROID	Thyroid condition diagnosed by physician	1/247	0.4	One thyroid condition was not coded and included in the analysis file.

Variable (SAS variable name from the analysis files)	Description of Variable	No. of Errors Found/No. of Questionnaires Examined	% Errors	Comments
THYRX	Thyroid medication (monthly consumption)	1/247	0.4	One subject was coded "95" when the questionnaire shows thyroid medication is taken 120 times per month.
TYLENOLB	Tylenolâ (in years)	0/247	0	
TYLENOLX	Tylenolâ (monthly consumption)	0/247	0	
WATER	Water (source of drinking water)	3/247	1.2	Bottled water not entered for two subjects. The analysis file contained a blank where city water should have been coded in another case.
WATERADD	Water additives used to soften drinking water	0/247	0	
WINEPR	Wine (previous amount of consumption)	1/247	0.4	One questionnaire showed some slight consumption of wine, but a dot appears in the analysis file.
WINEPRYR	Wine (years of previous wine consumption)	0/247	0	
YRSLIVE	Years resident in present neighborhood	0/247	0	

[1] Note that two questionnaires were missing and one copied questionnaire did not match the requested identification number.