ANALYSIS OF IMPACT OF MISSING DATA IN THE STUDY OF RACIAL DIFFERENCES IN ENDOMETRIAL CANCER SURVIVAL

by

Xinxin Dong

BMed, Beijing University, China, 2007

Submitted to the Graduate Faculty of

the Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Master of Science

University of Pittsburgh

2009

UNIVERSITY OF PITTSBURGH

Graduate School of Public Health

This thesis was presented

by

Xinxin Dong

It was defended on

April 20, 2009

and approved by

Thesis Advisor:
Carol K. Redmond, ScD
Distinguished Service Professor of Public Health
Department of Biostatistics
Graduate School of Public Health
University of Pittsburgh

Gong Tang, PhD
Assistant Professor
Department of Biostatistics
Graduate School of Public Health
University of Pittsburgh

Jeanette M. Trauth, MPA, MS, PhD
Associate Professor
Department of Behavioral and Community Health Sciences
Graduate School of Public Health
University of Pittsburgh

ANALYSIS OF IMPACT OF MISSING DATA IN THE STUDY OF RACIAL DIFFERENCES IN ENDOMETRIAL CANCER SURVIVAL

Xinxin Dong, M.S.

University of Pittsburgh, 2009

Endometrial cancer is the third most common cause of gynecologic cancer death and shows the largest overall survival difference (34%) between the races. The National Cancer Institute (NCI) Black/White Cancer Survival Study was a population-based study of racial differences in cancer survival. Endometrial cancer cases consisted of 149 black women, ages 20-79 years, residing in three selected metropolitan areas, who were diagnosed with endometrial cancer between 1985 and 1987. Cases were frequency matched in a ratio of approximately 1:2 to a sample of 341 white women with endometrial cancer. Information was derived from abstracts of hospital and physicians' records, centralized pathology review, and interviews. Potential explanatory factors for black-white survival differences have been previously investigated using Cox regression. However, there was a high proportion of missing values since 24 percent of patients were never interviewed. Some values were also missing for three other variables derived from medical records. Missing values may introduced bias in previous findings based only on the information available.

The primary objective of this thesis is to evaluate the effect of missing data on the estimated black/white mortality ratios adjusted for various explanatory factors. A second objective is to obtain more precise confidence intervals for the estimated mortality ratios. Nearest neighbor hot deck imputation has been used to generate fifty "complete" datasets. Adjusting for age and geographic location, the black/white mortality ratio for the imputed datasets was 3.3. When adjusted for all covariates, the mortality ratio was only 1.2. Overall, 87% of the excess mortality could be attributed to racial differences in disease stage, tumor characteristics, treatment, sociodemographic characteristics, hormonal and reproductive factors, the number of comorbidities and health behavior.

The results based on multiple imputation indicate that missing data did not introduce major bias in the earlier analyses. However, multiple imputation provided narrower confidence intervals than those obtained previously. Multiple imputation was worthwhile since it gave more precise estimates for the relative mortality ratios.

These findings have public health importance: they have implications for development of health policies and planning interventions to reduce the excess risk of death among black women with endometrial cancer.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

It is difficult to overstate my sincere gratitude to my thesis advisor, Dr. Carol K. Redmond. With her inspiration, patience, encouragement, and great efforts to explain things clearly and simply, she helped me go through all the difficult times and give me an invaluable research experience. Throughout my thesis-writing period, she extended abundant time and effect. This thesis would not be possible without her. Many thanks to Dr. Gong Tang for teaching me the techniques for dealing with missing data and sharing the code with me. He provided wise advice, many good ideas and kind assistance to my analysis. Working with him not only improved my statistical skills, but also increased my interest in missing data.I would like to thank Dr. Jeanette M. Trauth for carefully reading my thesis, picking up important errors and giving me helpful comments and suggestions. I appreciate all the thoughtful feedback from her.

I am grateful to all the administration officers in the Department of Biostatistics and in the Graduate School of Public Health. They assisted me in many ways in the past two years.

The original Black/White Cancer Survival Study was supported by the National Cancer Institute (N01-CN-25501; N01-CN-35043; N01-CN-45174; No1-45175; N01-CN-45176; N01-CN-75401). I also thank the National Cancer Institute Black/White Cancer Survival Study group for conducting this profound study and sharing these important data with me.

Finally, much gratitude is owed to my family, friends and classmates for all of their support and encouragement.

1.0 INTRODUCTION

Endometrial cancer refers to several types of malignancy, which arise from the endometrium or lining of the uterus. It is the most common gynecologic cancer in the United States. The most common subtype is endometrioid adenocarcinoma, which typically occurs within a few decades of menopause, and is associated with excessive estrogen exposure. This subtype often develops in the setting of endometrial hyperplasia, and presents most often with vaginal bleeding. It comprises about 75 to 80% of all endometrial cancers. The second most common form is papillary serous adenocarcinoma, which comprises about 10% of all endometrial cancers, and another form is clear cell adenocarcinoma (about 4-5% of all endometrial carcinomas). Both papillary serous and clear cell adenocarcinomas tend to be more aggressive than endometrioid adenocarcinomas, and are often detected at advanced stages. Sometimes an endometrial cancer has features of more than one subtype; this is called a mixed adenocarcinoma and they make up about 10% of all endometrial cancers. There are a few other rare types like mucinous adenocarcinoma and squamous cell adenocarcinoma that each compromise less than 1% of endometrial cancers (Dolinsky, 2008).

Endometrial carcinoma is the third most common cause of gynecologic cancer death (after ovarian and cervical cancer). On January 1, 2005, in the United States there were approximately 572,626 women alive who had a history of cancer of the corpus and uterus. In 2008, there are 40,100 new cases and 7,470 deaths from endometrial cancer in the United States (Ries et al., 2008).

There is a significant observed black/white differential in endometrial cancer survival. The incidence rate for African American women is 20.3 per 100,000 women and for Caucasian women is 24.3 per 100,000 women based on cases diagnosed in 2001-2005 from 17 SEER (Surveillance, Epidemiology, and End Results) geographic areas. However, the mortality rate is 3.9 per 100,000 white omen, and 7.1 per 100,000 black women, based on patients who died in 2001-2005 for the United States. Five-year relative survival rates for 1996-2004 from 17 SEER geographic areas were: 84.7% for white women and 61.1% for black women. For both races, 1.31% of women will develop endometrial cancer between their 50th and 70th birthdays based on the rate from 2003-2005. Decreases in mortality rates were observed for both races between 1975 and 2005, with the largest declines from 1975 through 1979 (Ries et al., 2008). However, survival remains much lower among black women of all ages and across all stages of disease (Miller et al., 1993).

1.1 THE NCI BLACK/WHITE CANCER SURVIVAL STUDY

The National Cancer Institute conducted a social-epidemiologic study of possible behavioral and biologic determinants of black/white racial disparities in cancer survival. This study selected four organ sites: cancers of the uterine corpus, breast, bladder, and colon (Howard et al., 1992). Cancer of the uterine corpus was chosen because it showed the largest overall racial survival difference (34%). Bladder cancer was selected because the survival difference was greater than 20% for each sex (Myers and Hankey, 1980). Female breast cancer was included because it was the leading cause of deaths attributable to cancer among women at that time (American Cancer Society, 1982), and the black/white difference in survival was greater than 10% after adjustment for age and stage (Myers and Hankey, 1980). Although the data from the SEER Program for the diagnostic period 1973 to 1979 indicated a much smaller racial differential (4%) in survival from colon cancer (Ries et al., 1983), it was retained since (1) it was the only site relevant to both males and females other than bladder cancer; (2) it was the second leading cause of deaths attributable to cancer; (3) a significant black/white difference in colon cancer survival (17%) was found in the SEER data from Atlanta (Myers and Hankey, 1984), which provided an opportunity to evaluate regional disparities in racial survival differences; and (4) racial differences in various behavioral and biologic factors could be correlated with the size of survival differences found across cancer sites (Howard et al., 1992).

Three population-based cancer registries were selected as data collection centers: the Georgia Center for Cancer Statistics, which covered the metropolitan Atlanta area; the Louisiana Tumor Registry, which covered the metropolitan New Orleans area; and the California Tumor Registry, which covered the metropolitan San Francisco/Oakland area. The absence of a North Central or Northeast area was a recognized limitation of the study.

All black women diagnosed with cancer of the four selected sites during the three years period, 1985 to 1987, were included in the study. Black cases were frequency matched to white cases on the basis of age group (< 50, 50-64 and 65 years of age), geographic area Atlanta, New Orleans, or San Francisco/Oakland) and year of diagnosis. For breast and colon cancer cases the frequency matching was in the ratio of 1:1, whereas for endometrial and bladder cancer, a ratio of 1:2 was chosen to increase the total sample size and provide more reliable estimates of the survival ratio comparing blacks to whites.

Four types of research instruments were used for data collection in this study: an abstract of the hospital record, a supplementary abstract of selected items from physician's office records, a pathology review of representative biopsy and surgical specimens, and a personal interview with the patient.

This study involved rapid reporting of patients newly diagnosed with cancer so that sampled persons could be interviewed as close to their initial diagnosis as possible. Patients were eligible for selection into the study if they met the following criteria:

1. had invasive or in situ cancer of the uterine corpus, bladder, colon, or female breast, excluding lymphomas of these sites;

2. had no prior cancer except basal or squamous cell carcinoma of the skin;

3. were either white or black as stated in the hospital record or by the patient;

4. were 20 to 79 years of age;

5. were resident at diagnosis in one of the geographic areas included in the study;

6. had a diagnosis of cancer between January 1, 1985, and December 31, 1986, for breast and colon, or through December 31, 1987, for bladder and uterine corpus.

7. The ability to speak English was added after two years of case accrual.

1.2 EXPLANATORY VARIABLES AND HYPOTHESES

Based on previous studies, potential explanatory variables included in the NCI Black/White Cancer Survival Study were factors characterizing stage at diagnosis, histopathologic characteristics of the tumor, treatment, concurrent medical conditions, health behavior items, hormonal and reproductive history variables, and sociodemographic factors.

The seven hypotheses guiding the design of this study were:

1. that black patients with cancer might report less pre-diagnostic screening activity, less recognition of cancer symptoms, fewer asymptomatic detections, longer duration of symptoms, and slower processing within the medical care system than white patients;

2. that blacks would be under-staged more frequently than whites based largely on speculation concerning the quality of medical evaluations for the two races;

3. that histological characteristics would account for some proportion of the black/white survival differences based on biological considerations relating to the cancers selected;

4. that host vulnerability (Dutton, 1979) involving three dimensions might contribute to the poorer survival of indigent patients or blacks. The three dimensions considered were poorer nutritional status, poorer overall health status, and alcoholism-related compromise of the immune system;

5. that black patients with cancer might be prescribed less aggressive treatment than white patients with comparable stages of disease;

6. that black patients for various reasons might have less adherence to the treatment prescribed; and

7. that social support and coping strategies might affect survival differences by facilitating adaptive behavior or compliance with treatment.

1.3 STUDY SAMPLE AND INCOMPLETENESS

At the end of 1992 there were 3380 eligible patients with histological or ascertained dates of diagnosis during the initial study period of January 1, 1985, through December 31, 1986 for breast and colon cancer, and January 1, 1985, through December 31, 1987 for bladder and endometrial cancer, which contained 649 blacks and 573 whites for breast cancer; 493 blacks and 574 whites for colon cancer; 194 blacks and 384 whites for bladder cancers; and 168 blacks and 353 whites for endometrial cancer. These patients were followed up from the date they were diagnosed through 2000.

The major reason for incomplete records in this study is the absence of interviews. The overall interview response rate for this study was 77%. The rates essentially were similar for both races, 76% for blacks and 77% for whites, but they varied by data collection center, 81% in San Francisco/Oakland, 76% in Atlanta, and 70% in New Orleans. The interview response rate also varied by organ site: 83% for breast, 71% for colon, and 75% for uterine corpus and bladder. The reasons for patient non-response were: (1) No attempt was made to contact the patient because physician consent was not obtained in 275 cases (8%); (2) When contact was attempted, 6% of the total study population had died or were too ill to be interviewed, 8% declined to be interviewed, and

2% were non-respondents for other reasons, such as distant residential moves or lost to follow-up. Other reasons for incompleteness are lack of permission to abstract hospital records and review microscopic slides and lost to follow-up.

1.4 PREVIOUS PUBLICATIONS

There are more than twenty-two published articles based on the NCI Black/White Cancer Survival Study, that have identified multiple factors associated with the black/white differences in survival for each of the four cancer sites. Since this thesis focuses on analyses of the black/white differences for survival for patients with endometrial cancer, we will review here only those papers that have dealt with that cancer site.

Hill et al. (1995) found over 75% of the excess of poorly differentiated tumors versus well-differentiated tumors among blacks could be explained by racial differences in use of replacement estrogens, age at first pregnancy, history of oophorectomy, poverty, stage of disease, use of screening, and access to health care. The most prominent factor was estrogen therapy, which was associated with favorable tumor grade and was used much less frequently by blacks. Barrett et al. (1995) indicated that high-grade (poorly differentiated) lesions increased the risk for stage III or IV disease (odds ratio 8.3, 95% confidence interval 3.4 to 20.3), as did serous histological subtype (odds ratio 3.5, 95% confidence interval 1.4 to 8.8) and no usual source of care (odds ratio 5.5, 95% confidence interval 1.4 to 20.9). In the final statistical model these three factors also accounted for the majority of the excess of advanced stage for blacks. Coates et al. (1996) found that time from symptom recognition to initial medical consultation does not contribute importantly to the more advanced stage of the endometrial cancer commonly found among black women, since: (1) median recalled times between symptom recognition and consultation were 16 days for black women and 14 days for white women; (2) the adjusted consultation rate among black women was only somewhat lower (0.87) than among white women, and the 95 percent confidence interval (0.58 to 1.31) was consistent with no true difference between the races; (3) the median time to consultation for women with stage IV cancer was only 15 days longer than the time (14 days) for the women with stage I cancer. Hill et al. (1996) found that adjusting for age and geographic location, risk of death among black women was 4.0 (95% confidence interval [CI] 2.8, 5.6) time that of white women. Approximately 40% of this difference could be attributed to a more advanced stage at diagnosis among black women, and 23% to tumor characteristics and treatment. Further adjustment for all remaining factors reduced the hazard ratio to 1.6 (95% CI 1.0, 2.6).

2.0 STATEMENT OF THE PROBLEM

Endometrial cancer was selected for this analysis because this cancer site had the largest overall racial survival difference (34 %). Other factors involved in the selection of endometrial cancer cases for further evaluation were that the interview response rate was relatively low (75%) for endometrial cancer, which means that the proportion of incomplete data is relatively high for most of the variables. In addition, there was a suggestion of a possible interaction among race, survival, and interview status. There were only four published articles based on the NCI Black/White Cancer Survival Study, that investigated the potential explanatory factors for the lower survival rate among black women with endometrial cancer when compared to white women: Hill et al. (1995); Barrett II et al. (1995); Coates et al. (1996); and Hill et al. (1996). All four papers derived their findings and conclusions based on the incomplete data. An "unknown" category was included for each variable that had at least ten subjects with missing data. If fewer than ten subjects were missing, those subjects were excluded from analyses using that variable.

However, there are 13 out of 21 variables having missing data (62%) with a missing proportion ranging from 5% to 43% for blacks and 11% to 38% for whites. Moreover, it appears that the non-responders have lower survival rates in comparison to responders. Thus, the conclusions derived in the papers might be biased.

3.0 STUDY OBJECTIVES

The main objective of this master’s thesis is to explore the impact of missing data by comparing the estimates of the black/white hazard ratio for mortality in Cox regression models utilizing the original updated dataset to those based upon imputation of the missing data. The specific aims of the imputation of missing data are to reduce possible bias introduced by the use of incomplete data and to achieve more reliable and precise findings for potential explanatory factors that account for black/white racial disparities in endometrial cancer survival. A secondary aim is to present analyses that are based on additional long-term follow-up that has been completed since the publication of the first four papers on black/white differences in endometrial cancer survival.

4.0 METHODS

4.1 DESCRIPTION OF THE ENDOMETRIAL CANCER DATASET

The updated dataset contains 490 women (149 black, 341 white), who have been followed up from the date of diagnosis (1985 -1987) through the year 2000. Among blacks 99 deaths have been observed and among whites 102 deaths have been observed. Potential explanatory variables studied for racial differences in survival are design variables (age and geographic location), stage at diagnosis based on the International Federation of Gynecology and Obstetrics (FIGO) staging system (FIGO, 1988), pathologic grade, comorbid conditions, symptoms at diagnosis, patient delay, total delay, smoking history, income level, insurance, usual source of care, poverty index, occupation class, education, marital status, body mass index quartile, histological subtype, oral contraceptive use, menopausal status, and treatment1. All these variables are categorical. Appendix A provides a list of the variables in the dataset along with the categories for each variable.

Table 1 shows the frequency distribution and 5 year survival rates of patients by race in each variable. Chi-square tests were carried out to test the equivalence of survival rates for black and white women for each categorical variable individually since the majority of the patients had complete follow-up (only 6 censored cases). The overall five-year survival rate in black women is 0.4748 (95% confidence interval [CI] 0.3944, 0.5552) and in white women is 0.8138 (95% CI 0.7722, 0.8554). The chi-square test shows a statistically significant overall racial difference in survival (chi-square 72.58, p value < .0001). This significant racial difference in survival also exists in the majority of subgroups defined by each variable. One hundred and nine (73.2%) black and two hundred and sixty-two (76.8%) white had been interviewed. The five-year survival rate in non-interviewed blacks is 0.2143 (95% CI 0.0902, 0.3384), and in the blacks interviewed the five-year survival is 0.5774 (95% CI 0.4835, 0.6713), which significantly differ from each other. For non-interviewed whites five-year survival is 0.7273 (95% CI 0.6277, 0.8269), and in interviewed whites it is 0.8392 (95% CI 0.7947, 0.8837), which shows the same trend as in blacks. However, there is not a significant difference between the interviewed and non-interviewed white women.

The racial difference in survival is significant in both interviewed and non-interviewed patients (chi-square 38.29 and 40.47, respectively, p values both < 0.0001). The racial difference between survival for interviewed and non-interviewed cancer patients is more pronounced for black than for white endometrial cancer patients, although a formal test for interaction between interview status and race on survival yields a p value of 0.11. Since many variables were derived from personal interviews, there is concern that bias may have been introduced into the multivariable analyses done in earlier publications, which were based on the dataset of patients with complete information on all variables.

The difference in stage at diagnosis, which was determined to be the most prominent factor in explaining the racial disparity in endometrial cancer survival rates in the paper by Hill et al. (1996), was not derived from personal interviews. It is interesting to evaluate whether the black-to-white hazard ratio differs by interview status within each stage since patients with more advanced stage had a lower interview rate. Table 2 shows that the black-to-white hazard ratio is higher in non-interviewed patients compared to interviewed patients in stage 1 and 2, but is lower in non-interviewed patients in stage 3 and 4. This suggests a 3-way interaction effect among race, stage and interview status.

4.2 COMPARISON OF RESULTS DERIVED FROM UPDATED DATASET FOR ALL CASES AND INTERVIEWED CASES

Cox proportional hazards models are used to explore the effect of each variable on survival time. The results are somewhat different when the models are built with original updated data for all cases and treat unknowns as a separate category compared to only interviewed cases, especially for those variables derived from personal interviews. The effects of symptoms, patient delay, total delay, and smoking history on survival are significant for cases with complete information, but become non-significant when only interviewed cases are analysed. Table 3 shows the results comparing the estimated hazard ratio for mortality between black and white patients for selected Cox proportional hazards models. The hazard ratios estimated based on all cases with complete information are a little higher than those estimated based on interviewed cases for each of the covariates individually, except when controlling for stage, they are similar to each other. This observation indicates that the racial difference in survival is smaller in interviewed women compared to non-interviewed women. The multivariable analyses, which were based on endometrial cancer cases with complete information, are inefficient and subject to bias.

4.3 IMPUTATION APPROACHES

4.3.1 Hot-Deck Imputation

Imputation is a very popular technique for dealing with the missing-data problem. It involves filling in the missing values under a certain prediction model and then analyzing this imputed dataset as if there were no missing observations. As one of the often-used imputation methods, hot deck imputation is defined as a method where an imputed value is selected from an estimated distribution for each missing value. One approach is the nearest neighbor hot deck, which is to define a metric to measure distance between units, based on the values of covariates, and then to choose imputed values that come from responding units close to the unit with the missing value. For example, let x_{i1}, …, x_{ij} be the values of J appropriately scaled covariates for a unit i for which y_i is missing. Define the distance between units i and i' as

$$d\ (i, i') = \max |\ x_{ij} - x_{i'j}|$$

We might choose an imputed value for y_i from those unit i' that are such that (1) $y_{i'}$, $x_{i'1,}$ …, $x_{i'j}$ are observed, and (2) d (i, i') is less than some value d_0. The number of candidates i' can be controlled by varying the value of d_0 .(Little and Rubin, 1986)

In the endometrial cancer dataset the unit is an individual case with its corresponding values for the covariates and survival outcome. The procedure consists of three stages. First, a computer program is written for the imputation to measure the distance between units based on the covariates whose values are known and the survival outcome since it is closely related to all the variables that have missing values. Second, the technique identifies the subset of units with known responses that has the smallest distance from a unit with a missing value. Third, a value is randomly drawn from the subset of units with known responses to insert in place of the missing value.

Multiple-imputation is an extension of a single imputation where the imputation process described above is repeated *n* times. For each of the *n* independent imputations, a "complete" dataset is created in which an imputed value is substituted for each missing value. These *n* datasets are analyzed separately and the results are combined to form one overall inference. Multiple imputation overcomes one important limitation of single imputation, which is that standard variance formulas applied to the imputed values for missing data systematically underestimate the variance of estimates, even when the model used to generate the imputations is correct. (Little and Rubin, 1986)

4.3.2 Preparation of endometrial cancer dataset for multiple imputation

In the NCI Black/White Cancer Survival study, all the variables that need to be imputed for the endometrial cancer cases have a significant effect on survival time (Table 3). During the first five years of follow-up, information on follow-up for survival is nearly complete, with only 6 cases for which follow-up time is censored (cases lost-to-follow-up) prior to five years (Table 4). Therefore, survival time can be used for determining the distances between units. First, a categorical variable, which represents survival time, has been created, in which the 6 censored survival times are assumed to have survived at least

5 years (Table 5). Among the covariates, there are five variables that have fewer than five missing values for either race (table 6).

Both stage and pathologic grade are highly significantly correlated with survival time. Based an distribution of survival time for cases with known stage at diagnosis, the one missing value for stage has been imputed as Stage 1 since its survival time is between one and three years (Table 7 & 8). For the three patients with missing pathologic grades the missing value is imputed as Grade 3 for two cases whose survival time was less than five years, and Grade 1 for the one patient who survived more than five years (Table 9 & 10). Symptoms and patient delay have a direct relationship in that there is no patient delay if there are no symptoms at the time of the first hospital visit. Therefore, there is no patient delay for the two cases with missing values for symptoms and the missing values are imputed as no symptoms. Menopausal status is closely associated with patient age. Therefore, menopausal status has been imputed as premenopausal for patient 20-49 years old and postmenopausal for those patients 50-64 years old (Table 11).

After insertion of values for the five variables above for cases with missing information they were treated along with design variables (age and geographic location), race and survival time as the variables with complete information used for generating imputation code to determine the distance between units for imputing other variables with missing values.

There are three reasons for incompleteness in the data: 1) Lack of permission to abstract of hospital records or microscopic slides leads to incomplete records for variables derived from medical records, physician records, and pathology review; 2) Absence of interview causes variables derived from the interview to be missing; 3) Non-response to some of the interview questions make variables derived from the interview incomplete even for patients who have interviews.

1. V1 is defined as variables missing because of lack of access to medical records, physician records and pathology review, including comorbidities, BMI and therapy (Table 12).

2. V2 is defined as variables missing only because of absence of interview, including smoking history, insurance, education, oral contraceptive use, usual source of care and occupation class (Table 13).

3. V3 is defined as variables missing both because of absence of interview and non-response in the interview, including income level, poverty index, total delay and patient delay (Table 14).

Other data preparation tasks included: 1) To avoid sparse data, poverty index group 4 (301-400) and 5 (>400) were combined into one group (>300) (Please see Appendix A for listing and explanation of variables in the dataset); 2) In order to provide a more natural order for occupational class that corresponded with socioeconomic variables, the codes for the categories were reassigned as shown in Appendix A.

4.3.3 Generation of the imputation code-Illustrative example (Poverty Index)

The poverty index is an explanatory variable of major interest in the study of factors related to differential mortality between black and white patients with endometrial cancer. In addition, the poverty index, which is computed from the income level and household size, is missing for 37% of cases (41% for blacks and 32% for whites). Not only is it missing due to the absence of an interview, but also there are patients interviewed who do not provide information on income, which is needed to calculate the index. During the first stage of the imputation process, we utilize contingency table analysis to identify other variables closely associated with the poverty index. The contingency coefficient, which was the measure of association chosen for the analysis, was calculated for each cross-classification of the poverty index with other variables. The variables were then ranked according to the value of the contingency coefficient of their association with poverty index (Table 15). The second stage of the imputation process involved fitting log-linear models to find a subset of multiple variables significantly correlated with poverty index (Table 16) from which the missing values of the poverty index could then be imputed.

Race, location, age group and occupational class are four variables highly correlated with the poverty index, but they cannot be used to impute all the missing values for the poverty index since occupational class is also missing for 119 cases (24%) that do not have interviews. Table 17 shows the distribution of the poverty index by occupational class according to whether the data was present or missing.

At the third stage we used these four variables (race, location, age group and occupational class) along with survival time to construct the imputation code for the 51 (10%) of interviewed cases who were missing information on the poverty index, but who had a known occupational class. The specific code for the imputation was:

povgpcode= survival time * 10000 + race * 1000 + location * 100 + occupation class * 10+ age group * 1

For the 119 (24%) of cases that did not have interview data, race, location, age group and survival time was used in the imputation code, since occupational class was not available. The code for this imputation was:

povgpcode2= Survival time*1000+Race*100+Location*10+Age group*1

Analogous procedures were carried out to impute values for other variables with missing data. (Appendix B).

4.3.4 Multiple imputation

Multiple imputations were carried out to overcome an obvious disadvantage of single imputation, that it cannot reflect sampling variability under one model for non-response or uncertainty about the correct model for non-response. With multiple imputations, the resulting n complete-data analyses can be easily combined to create an inference that valid reflects sampling variability because of the missing value, and uncertainty about the correct model was displayed by the variation in valid inferences across the models since more than one model was used.

R (a programming language and software environment for statistical computing and graphics) was used to carry out the nearest neighbor hot deck imputation process. First, a subset of subjects with complete information, who have the closest distance to the case with a missing value was identified by comparing the difference in imputation code between the cases with complete information with the case whose information was missing. Then a value was drawn at random from the subset of cases with known values to replace the missing values. This procedure was repeated 50 times to get 50 new "complete" datasets. The code for this stage is provided in Appendix C.

The Cox regression analysis for each of the 50 imputed datasets was carried out in order to obtain more reliable estimates and the variability associated with the imputation of the missing values. The eight steps in the multiple imputation are:

1. Build the models with selected covariates in each of the 50 imputed datasets.

2. Calculate the estimated coefficients (Q_d, d=1,…,50) and standard errors (W_d, d=1,…,50) for race in the Cox proportional hazards models.

3. Estimate the combined coefficient for race in each model by

$$\overline{\theta}_{50} = \frac{1}{50}\sum_{d=1}^{50}\overline{\theta_d} \;;$$

4. Estimate the combined black to white hazard ratio in each model by

$$HR_{race} = \exp(\overline{\theta}_{50});$$

5. Estimate the average within-imputation variance by

$$\overline{W}_{50} = \frac{1}{50}\sum_{d=1}^{50}\overline{W_d} \;;$$

6. Estimate the between-imputation variance by

$$B_{50} = \frac{1}{50-1}\sum_{d-1}^{50}(\hat{\theta}_d - \overline{\theta}_{50})^2 \;;$$

7. Estimate the total variability associated with the imputation of the missing values by

$$T_{50} = \overline{W}_{50} + \frac{50+1}{50}B_{50} \;;$$

8. At the final step estimate the fraction of information about the black to white survival ratio missing due to incomplete data by

$$\hat{\gamma}_{50} = (1+1/50)B_{50}/T_{50} \,.$$

5.0 RESULTS

5.1 DISTRIBUTION OF IMPUTED DATA COMPARED TO ORIGINAL DATA- ILLUSTRATIVE EXAMPLE FOR THE POVERTY INDEX

In order to verify that the imputation has been carried out appropriately, the properties of the dataset from one of the imputations is examined. As would be expected, there has been no obvious change in the frequency distributions (Table 18, Figure 1 and Figure 2) or hazard ratios for mortality (Table 19) when the original updated data on the poverty index are compared to the imputed dataset. Similar comparisons of other variables in original and imputed datasets are provided in Appendix D.

5.2 MULTIVARIABLE COX REGRESSION MODELS OF THE BLACK/WHITE HAZARD RATIOS ADJUSTING FOR SELECTED COVARIATES BASED ON THE ORIGINAL DATASET

The second line for each Cox regression model with selected variables in Table 20 shows the estimates of hazard ratios for mortality among black women compared to white women with endometrial cancer after controlling for selected covariates utilizing the original updated dataset. Adjustment for sociodemographic factors (marital status, poverty index, occupation, usual source of care, insurance, education) has a dramatic impact on the racial difference in survival from endometrial cancer, while hormonal and reproductive factors (menopausal status, use of oral contraceptives) and treatment have a relatively small effect. Stage, tumor characteristics (pathologic grade, histological tumor subtype), comorbidities and health behavior (number of comorbid conditions, body mass index (BMI), smoking history, total delay, patient delay) have a moderate influence on the racial survival differences. After multivariable adjustment for all covariates, the racial hazards ratio becomes 1.3 (95% confidence interval 0.7, 2.6).

When we use a stepwise procedure, setting the p value for entering at 0.20, and the p value for removing a covariate at 0.10 in order to select covariates for the full model, stage, poverty index, histology, treatment, type of insurance, patient delay, occupational class, BMI, usual source of care, and total delay would be included. The hazards ratio for mortality among black women compared to white women estimated in this full model is 1.8 with a 95% confidence interval of 1.0 to 3.1. Therefore, stage, poverty index, histology, treatment, type of insurance, patient delay, occupational class, BMI, usual source of care, and total delay explain 77% of the reduction in the black to white hazard ratio for mortality by including all covariates.

5.3 MULTIVARIABLE COX REGRESSION MODELS OF THE BLACK/WHITE HAZARD RATIOS ADJUSTING FOR SELECTED COVARIATES BASED ON THE MULTIPLE IMPUTED DATASETS

The third line for each Cox regression model with selected variables in Table 20 shows the combined estimates of hazard ratios for mortality among black women compared to white women after controlling for selected covariates utilizing the 50 imputed datasets. Sociodemographic factors (marital status, poverty index, occupation, usual source of care, type of insurance, education) still had a dramatic impact on the racial difference in survival from endometrial cancer. The hormonal and reproductive factors (menopausal status, use of oral contraceptives) have the smallest effect. In multivariable adjustment for all covariates, the hazard ratio of black women compared to white women is 1.2, which is a little less than the hazard ratio estimated from the original data. In addition, the 95% confidence intervals for all the estimated hazards ratios are narrower than those estimated with the original data.

The estimate of average within-imputation variance is 0.048, the between-imputation variance is 0.0049, and the total variability associated with the imputation of the missing values is 0.053. Overall, 9.4% of the information of explanatory variables related to the black/white hazard ratio is missing.

6.0 DISCUSSION

The previous analyses by Hill (1996) showed that the risk of death among black women was 4.0 times greater (95% confidence interval [CI] 2.8, 5.6) than that of white women when only the design variables (age and geographic location) are taken into account. Approximately 40% of this difference could be attributed to a more advanced stage at diagnosis among black women, and 23% to tumor characteristics and treatment. Adjustment for all remaining factors reduced the hazard ratio further to 1.6 (95% CI 1.0, 2.6).

With the updated follow-up data the black/white hazard ratio for mortality was 3.3 after adjustment for age and geographic location and 1.3 after adjustment for all covariates. Although the hazard ratio for mortality decreased somewhat with long-term follow-up, it is still substantially elevated. As in the prior analyses, the most important factor related to black/white differences was stage of disease at diagnosis. More advanced stage at diagnosis among black women explained 35% of the excess mortality among black women. The reason for the somewhat lower overall black/white hazard ratio for mortality in the present analyses (3.3 versus 4.0) may be that with longer observation time, there is a tendency for the early marked differences to become attenuated. In other words the observed black/white survival difference is less pronounced in additional follow-up among those women who survived longer compared to those who died within the first few years following diagnosis.

The combined estimates of hazard ratios of mortality for race from Cox multivariable proportional hazard models with selected covariates derived from the 50 imputed datasets are very similar to those estimated with the original updated dataset. Therefore, it appears that missing data did not introduce substantial bias into the analyses based on the original updated dataset, even though the fraction of missing data was as high as 9 to 10 percent for some of the covariates of interest. However, the 95% confidence intervals are much narrower with imputed data, which indicates that the imputation of the missing data increases the precision of the estimates.

Our analyses based on imputing missing observations continues to provide evidence that sociodemographic factors (marital status, poverty index, occupation, usual source of care, insurance, education) are the most important variables in explaining the disparity in endometrial cancer survival rates between black and white women. Taken together, they explain 65% of the excess mortality among black women after adjustment for age and geographical location (adjusted hazard ratio 1.8 versus unadjusted hazard ratio 3.3). Comorbidities and health behavior factors (comorbidity, BMI, smoking history, total delay, patient delay) also contribute substantially to the excess mortality (adjusted hazard ratio 2.5 versus unadjusted hazard ratio 3.3). Hormonal and reproductive factors (menopausal status, use of oral contraceptives) only explain a relative small proportion of the racial disparities in survival (adjusted hazard ratio 3.1 versus unadjusted hazard ratio 3.3). The lack of importance of menopausal status in the model may be related to its collinearity with age, which is included in all models since black women were frequency matched to white women by age.

There are several potential limitations to the current analysis:

All underlying prognostic factors may not have been included in the study. A potential prognostic variable absent in this analysis is peritoneal cytology.

Misclassification may have occurred in some of the covariates, such as those from the interview, which relate to sensitive information, such as income. Misclassification could also be present if black women are under-staged because they receive fewer staging tests than whites.

The imputation process may also have introduced some bias. When generating the imputation code for each variable, race was generally closely associated with the covariate of interest. Therefore, the imputations are always performed within a racial group.

7.0 CONCLUSION

Missing data did not introduce significant bias in the earlier papers since the estimates of the hazard ratio for mortality based on fifty imputed datasets were similar to those based on the updated original dataset. However, the multiple imputation provided estimates of confidence intervals that were narrower than those previously published. The multiple imputation was, therefore, worthwhile since it gave more precise estimates for the relative survival ratios. The most important explanatory variables for a lower survival rate among black women with endometrial cancer when compared to white women are sociodemographic factors (marital status, poverty index, occupation, usual source of care, insurance, education), stage at diagnosis, and comorbidities and health behavior factors (comorbidity, BMI, smoking history, total delay, patient delay). Overall, 87% of the excess mortality could be attributed to racial differences in stage at diagnosis, tumor characteristics, treatment, sociodemographic characteristics, hormonal and reproductive factors and factors related to number of comorbidities and health behavior.

8.0 TABLES AND FIGURES

Table 1: Five-year survival and its 95% confidence interval (CI) by race for each of the covariates

Study Variable	Category	Black				White				Chi-square test
		N	Deaths	5 year survival	95% CI	N	Deaths	5 year survival	95% CI	p-value*
Race		149	99	0.47	(0.39, 0.56)	341	102	0.81	(0.77, 0.86)	<.0001
Location$^+$										
	Atlanta	47	29	0.49	(0.34, 0.63)	96	26	0.85	(0.78, 0.92)	<.0001
	New Orleans	52	40	0.42	(0.29, 0.56)	90	29	0.79	(0.70, 0.87)	<.0001
	San Fran/Oak**	50	30	0.52	(0.38, 0.66)	155	47	0.80	(0.74, 0.87)	<.0001
Age Group$^+$										
	20-49	17	7	0.82	(0.64, 1.00)	44	4	0.90	(0.81, 0.99)	0.0037
	50-64	53	27	0.62	(0.49, 0.75)	120	27	0.87	(0.81, 0.93)	0.0001
	65-79	79	65	0.30	(0.20, 0.41)	177	71	0.76	(0.69, 0.82)	<.0001
Stage										
	Unknown	1	1	0.00	(0.00, 0.00)	0	0			
	1	96	52	0.66	(0.57, 0.76)	283	67	0.89	(0.85, 0.92)	<.0001
	2	14	12	0.21	(0.00, 0.43)	23	8	0.74	(0.56, 0.92)	0.0011
	3	18	15	0.17	(0.00, 0.34)	25	17	0.38	(0.18, 0.57)	0.17
	4	20	19	0.05	(0.00, 0.15)	10	10	0.00	(0.00, 0.00)	0.61
Interviewed										
	No	40	35	0.21	(0.09, 0.34)	79	34	0.73	(0.63, 0.83)	<.0001
	Yes	109	64	0.58	(0.48, 0.67)	262	68	0.84	(0.79, 0.88)	<.0001
Grade										
	Unknown	2	2	0.00	(0.00, 0.00)	1	0	1.00	(1.00, 1.00)	0.23
	1	51	22	0.70	(0.58, 0.83)	176	38	0.89	(0.85, 0.94)	0.0009
	2	51	38	0.45	(0.31, 0.59)	110	36	0.81	(0.74, 0.89)	<.0001
	3	45	37	0.27	(0.14, 0.40)	54	28	0.56	(0.42, 0.69)	0.0011
Comorbidity										
	Unknown	8	3	0.75	(0.45, 1.05)	37	9	0.84	(0.72, 0.96)	0.41
	No	21	12	0.57	(0.36, 0.78)	105	20	0.89	(0.83, 0.95)	<.0001
	Yes	120	84	0.44	(0.35, 0.53)	199	73	0.77	(0.71, 0.83)	<.0001

$_+$ Design variables

$*$ P value from chi-square test for the null hypothesis that the five-year survival rates are the same between black and white women in each of the category

$_{**}$ San Francisco-Oakland

Table 1: Five-year survival and its 95% confidence interval (CI) by race for each of the covariates (continued)

Study Variable	Category	Black				White				Chi-square test
		N	Deaths	5 year survival	95% CI	N	Deaths	5 year survival	95% CI	p-value*
Symptoms										
	Unknown	.	.	.		2	2	0.50	(0.00, 1.19)	.
	No	3	2	0.33	(0.00, 0.87)	20	3	0.95	(0.85, 1.05)	0.018
	Yes	146	97	0.48	(0.40, 0.56)	319	97	0.81	(0.76, 0.85)	<.0001
Patient Delay										
	Unknown	46	39	0.26	(0.13, 0.39)	86	35	0.74	(0.65, 0.84)	<.0001
	No	3	2	0.33	(0.00, 0.87)	22	5	0.91	(0.79, 1.03)	0.066
	0-<1m**	59	34	0.56	(0.43, 0.69)	155	46	0.82	(0.76, 0.88)	<.0001
	1-<3m**	15	8	0.60	(0.35, 0.85)	33	8	0.82	(0.69, 0.95)	0.035
	3-<6m**	10	5	0.70	(0.42, 0.98)	19	3	0.84	(0.67, 1.01)	0.12
	>=6m**	16	11	0.56	(0.31, 0.80)	26	5	0.92	(0.82, 1.03)	0.0002
Total Delay										
	Unknown	50	42	0.26	(0.14, 0.38)	94	39	0.74	(0.66, 0.83)	<.0001
	0-<1m**	39	20	0.64	(0.49, 0.79)	89	30	0.81	(0.73, 0.89)	0.066
	1-<3m**	22	19	0.32	(0.12, 0.51)	55	14	0.85	(0.76, 0.95)	<.0001
	3-<6m**	14	6	0.64	(0.39, 0.89)	38	7	0.84	(0.72, 0.96)	0.074
	>=6m**	24	12	0.71	(0.52, 0.89)	65	12	0.88	(0.80, 0.96)	0.0017
Smoking History										
	Unknown	42	36	0.21	(0.09, 0.34)	77	35	0.73	(0.63, 0.83)	<.0001
	Never	71	41	0.52	(0.40, 0.63)	145	38	0.85	(0.79, 0.91)	<.0001
	Former	25	16	0.64	(0.45, 0.83)	89	23	0.82	(0.74, 0.90)	0.0006
	Current	11	6	0.82	(0.59, 1.05)	30	6	0.86	(0.74, 0.99)	0.064
Income										
	<10K	60	41	0.49	(0.37, 0.62)	52	16	0.76	(0.64, 0.88)	0.0003
	10-<20K	19	9	0.68	(0.48, 0.89)	55	18	0.78	(0.67, 0.89)	0.24
	20-<35K	5	1	1.00	(1.00, 1.00)	64	13	0.89	(0.81, 0.97)	0.91
	35K	4	1	0.75	(0.33, 1.17)	64	8	0.95	(0.90, 1.00)	0.36
	Unknown	61	47	0.33	(0.21, 0.45)	106	47	0.73	(0.64, 0.81)	<.0001

** Month

* P value from chi-square test for the null hypothesis that the five-year survival rates are the same between black and white women in each of the category

Table 1: Five-year survival and its 95% confidence interval (CI) by race for each of the covariates (continued)

Study Variable	Category	Black				White				Chi-square test
		N	Deaths	5 year survival	95% CI	N	Deaths	5 year survival	95% CI	p-value*
Insurance										
	None	15	7	0.73	(0.51, 0.96)	12	5	0.67	(0.40, 0.93)	0.98
	Public	42	34	0.38	(0.23, 0.53)	25	7	0.83	(0.68, 0.98)	<.0001
	Any	50	22	0.70	(0.57, 0.83)	227	55	0.85	(0.80, 0.90)	0.0023
	Unknown	42	36	0.21	(0.09, 0.34)	77	35	0.73	(0.63, 0.83)	<.0001
Usual Source of Care										
	None	14	11	0.36	(0.11, 0.61)	27	5	0.84	(0.70, 0.98)	<.0001
	Public	32	21	0.63	(0.46, 0.79)	7	1	0.86	(0.60, 1.12)	0.047
	Private	61	31	0.60	(0.48, 0.73)	229	61	0.84	(0.79, 0.89)	<.0001
	Unknown	42	36	0.21	(0.09, 0.34)	78	35	0.73	(0.63, 0.83)	<.0001
Poverty Index										
	0-125	58	40	0.49	(0.37, 0.62)	46	14	0.75	(0.62, 0.88)	0.0006
	126-200	11	7	0.55	(0.25, 0.84)	22	5	0.86	(0.72, 1.01)	0.021
	201-300	12	4	0.83	(0.62, 1.04)	32	15	0.69	(0.53, 0.85)	0.40
	301-400	1	0	1.00	(1.00, 1.00)	32	6	0.88	(0.76, 0.99)	0.64
	>400	6	1	0.83	(0.54, 1.13)	100	15	0.94	(0.89, 0.99)	0.85
	Unknown	61	47	0.33	(0.21, 0.45)	109	47	0.73	(0.65, 0.82)	<.0001
Occupation Class										
	Unknown	43	37	0.23	(0.11, 0.36)	79	36	0.73	(0.64, 0.83)	<.0001
	Home-maker	15	8	0.60	(0.35, 0.85)	36	11	0.78	(0.64, 0.91)	0.21
	Mgt/Prof **	12	6	0.58	(0.30, 0.86)	75	16	0.85	(0.77, 0.93)	0.011
	Tech/Sales+	12	4	0.75	(0.51, 1.00)	97	19	0.91	(0.85, 0.96)	0.27
	Skilled	37	20	0.62	(0.46, 0.77)	38	13	0.75	(0.61, 0.89)	0.13
	Unskilled	30	24	0.43	(0.26, 0.61)	16	7	0.69	(0.46, 0.91)	0.032

** Management or professional worker

+ Technique or sales

* P value from chi-square test for the null hypothesis that the five-year survival rates are the same between black and white women in each of the category

Table 1: Five-year survival and its 95% confidence interval (CI) by race for each of the covariates (continued)

Study Variable	Category	Black				White				Chi-square test
		N	Deaths	5 year survival	95% CI	N	Deaths	5 year survival	95% CI	p-value*
Education										
	Unknown	42	36	0.21	(0.09, 0.34)	77	35	0.73	(0.63, 0.83)	<.0001
	< High school	59	39	0.52	(0.39, 0.65)	51	18	0.76	(0.65, 0.88)	0.0022
	High School Grad	29	14	0.69	(0.52, 0.86)	92	24	0.82	(0.74, 0.89)	0.030
	> High school	19	10	0.58	(0.36, 0.80)	121	25	0.89	(0.83, 0.95)	0.0004
Marital Status										
	Unknown	6	4	0.33	(0.00, 0.71)	2	0	1.00	(1.00, 1.00)	0.17
	Partnered	48	29	0.56	(0.42, 0.70)	181	43	0.86	(0.81, 0.91)	<.0001
	Widowed	63	45	0.41	(0.29, 0.53)	93	40	0.71	(0.62, 0.80)	0.0005
	Div/Sep**	26	19	0.46	(0.27, 0.65)	41	8	0.87	(0.77, 0.98)	<.0001
	Never married	6	2	0.67	(0.29, 1.04)	24	11	0.74	(0.56, 0.92)	0.77
Body Mass Index Quartile (BMI)										
	Unknown	32	28	0.19	(0.05, 0.32)	44	22	0.73	(0.60, 0.86)	<.0001
	Low normal	7	4	0.71	(0.38, 1.05)	99	22	0.90	(0.84, 0.96)	0.042
	High normal	11	5	0.64	(0.35, 0.92)	69	22	0.77	(0.67, 0.87)	0.33
	Overweight	30	19	0.42	(0.24, 0.60)	66	14	0.89	(0.82, 0.97)	<.0001
	Very overweight	69	43	0.58	(0.46, 0.70)	63	22	0.71	(0.60, 0.83)	0.0055
Histology										
	Others	45	35	0.32	(0.19, 0.46)	33	19	0.56	(0.39, 0.74)	0.058
	Adenosqua-mous	104	64	0.54	(0.44, 0.63)	308	83	0.84	(0.80, 0.88)	<.0001

** Divorced or separated

* P value from chi-square test for the null hypothesis that the five-year survival rates are the same between black and white women in each of the category

Table 1: Five-year survival and its 95% confidence interval (CI) by race for each of the covariates (continued)

Study Variable	Category	Black				White				Chi-square test
		N	Deaths	5 year survival	95% CI	N	Deaths	5 year survival	95% CI	p-value*
Treatment+										
	Surgery Only	42	18	0.74	(0.60, 0.87)	148	28	0.89	(0.84, 0.94)	0.0009
	No Surgery	39	36	0.15	(0.04, 0.27)	16	9	0.63	(0.39, 0.86)	0.0022
	Surg. Plus Chemo++	11	8	0.36	(0.08, 0.65)	12	9	0.33	(0.07, 0.60)	0.71
	Surg. Plus Hormn+++	6	5	0.17	(0.00, 0.46)	7	6	0.43	(0.06, 0.80)	0.87
	Surg Plus CandH++++	1	1	0.00	(0.00, 0.00)	5	5	0.20	(0.00, 0.55)	0.65
	Surg Plus RT**	48	31	0.50	(0.36, 0.64)	124	46	0.80	(0.73, 0.87)	0.0003
	Surg+ RT+ CT/HT***	16	12	0.31	(0.09, 0.54)	14	10	0.43	(0.17, 0.69)	0.0003
	Unknown	64	48	0.41	(0.29, 0.53)	129	49	0.77	(0.69, 0.84)	<.0001
Oral Contraceptive Use										
	No interview	46	36	0.21	(0.09, 0.34)	77	35	0.73	(0.63, 0.83)	<.0001
	Not used	94	60	0.53	(0.43, 0.63)	197	59	0.81	(0.44, 1.18)	<.0001
	Used	13	3	0.92	(0.78, 1.07)	67	8	0.92	(0.86, 0.99)	0.29
Menopausal status										
	Unknown	3	2	0.33	(0.00, 0.87)	3	0	1.00	(1.00, 1.00)	0.12
	Pre-menopausal	20	7	0.90	(0.77, 1.03)	59	6	0.93	(0.86, 1.00)	0.0096
	Post-menopausal	126	90	0.41	(0.33, 0.50)	279	96	0.79	(0.74, 0.84)	<.0001

+ Not mutually exclusive

++ Surgery plus chemotherapy; +++Surgery plus hormonotherapy, ++++ Surgery plus chemotherapy and hormonotherapy

Surgery plus radiotherapy; * Surgery plus radiotherapy and either chemotherapy or hormonotherapy

* P value from chi-square test for the null hypothesis that the five-year survival rates are the same between black and white women in each of the category

Table 2: Black to white hazard ratio for overall mortality and its 95% confidence interval (CI) by interview status and FICO stage

Stage	Interviewed	N		HR	95% CI
		White	Black		
1	No	67	23	4.01	(2.02, 7.98)
	Yes	216	73	3.03	(1.89, 4.86)
2	No	4	3	-*	-*
	Yes	19	11	4.70	(1.37, 16.13)
3	No	3	4	1.62	(0.14, 18.31)
	Yes	22	14	2.34	(0.78, 7.03)
4	No	3	12	0.00	-*
	Yes	7	8	1.66	(0.43, 6.39)

* Can not estimate because of inadequate sample size

Table 3: Black to white hazard ratio for overall mortality and its 95% confidence interval (CI) in selected Cox proportional hazards models

	In the dataset based on all cases (treat unknowns as a separate category)		In the dataset based on cases interviewed	
Variable other than race in the model:	Hazard Ratio	95% CI	Hazard Ratio	95% CI
None	3.27	(2.46, 4.34)	3.00	(2.11, 4.27)
Stage	2.52	(1.86, 3.40)	2.65	(1.84, 3.82)
Grade	2.83	(2.12, 3.77)	2.63	(1.85, 3.76)
Comorbidity	2.99	(2.24, 4.00)	2.80	(1.96, 4.00)
Symptoms	3.27	(2.46, 4.35)	2.97	(2.08, 4.23)
Patient Delay	3.45	(2.58, 4.60)	2.99	(2.09, 4.27)
Total Delay	3.48	(2.61, 4.63)	3.07	(2.15, 4.37)
Smoking history	3.54	(2.66, 4.72)	2.99	(2.10, 4.26)
Income	2.92	(2.15, 3.97)	2.36	(1.61, 3.48)
Insurance	3.28	(2.41, 4.47)	2.59	(1.75, 3.84)
Usual Source of Care	3.55	(2.63, 4.79)	2.94	(2.01, 4.28)
Poverty Index	2.84	(2.09, 3.86)	2.25	(1.53, 3.32)
Occupation Class	3.09	(2.26, 4.22)	2.30	(1.54, 3.43)
Education	3.36	(2.48, 4.56)	2.70	(1.84, 3.96)
Marital Status	3.08	(2.30, 4.13)	2.76	(1.92, 3.96)

Table 4: Censoring status by survival time

Censoring	Survival time			
Status	≤1 year	1<years≤3	3<years≤5	>5 years
Yes	2(3.33%)	3(5.56%)	1(3.03%)	283(82.51%)
No	58(96.67%)	51(94.44%)	32(96.97%)	60(17.49%)

Table 5: Define categorical survival time

Original Survival Time (years)	Survival Time Category
≤1 year	1
1<years≤3	2
3<years≤5	3
>5 years*	4

* assume those censored patients will survival greater than 5 years

Table 6: Complete covariates used to generate imputation code

Variable name	Black		White	
	Total	Missing #	Total	Missing #
Stage	149	1	341	0
Pathologic Grade	149	2	341	1
Symptoms	149	0	341	2
Histology	149	0	341	0
Menopausal status	149	3	341	3

Table 7: Distribution of stage by survival time

Stage	Survival Time (years)				
	≤1 years	1<years≤3	3<years≤5	>5 years*	Total
1	22	26	21	310	379
2	7	5	5	20	37
3	9	17	5	12	43
4	22	5	2	1	30
Total	60	53	33	343	489

* assume those censored patients will survival greater than 5 years

Table 8: Distribution of missing in stage by survival time

Stage missing	Survival Time (years)				
	≤1 year	1<years≤3	3<years≤5	>5 years*	Total
No	60	53	33	343	489
Yes	0	1	0	0	1
	60	54	33	343	490

* assume those censored patients will survival greater than 5 years

Table 9: Distribution of grade by survival time

Grade	Survival Time (years)				
	≤1 year	1<years≤3	3<years≤5	>5 years*	Total
1	10	16	10	191	227
2	20	23	9	109	161
3	29	15	13	42	99
Total	59	54	32	342	487

* assume those censored patients will survival greater than 5 years

Table 10 : Distribution of missing grade by survival time

Grade Missing	Survival Time (years)				
	≤1 year	1<years≤3	3<years≤5	>5 years*	Total
No	59	54	32	342	487
Yes	1	0	1	1	3
Total	60	54	33	343	490

* assume those censored patients will survival greater than 5 years

Table 11: Distribution of menopausal status by age group

Menopausal status	Age group			
	20-49	50-64	65-79	Total
Premenopausal	59	169	256	484
Postmenopausal	2	4	0	6
Total	61	173	256	490

Table 12 : Variables that are derived from medical records, physician records, and pathology review

	Black (n=149)	White (n=341)
Variable name	# Missing	# Missing
Comorbidity	8	37
BMI	32	44
Therapy	64	129

Table 13: Variables missing only because of non-interview

	Black (n=149)		White (n=341)	
Variable name	# Non-interviewed	# Missing	# Non-interviewed	# Missing
Smoking History	42	42	77	77
Insurance	42	42	77	77
Education	42	42	77	77
Oral Contraceptive Use	42	42	77	77
Usual Source of Care	42	42	77	78
Occupation Class	42	43	77	79

Table 14: Variables missing because of non-interview and non-response in the interview

	Black (n=149)		White (n=341)	
Variable name	# Non-interviewed	# Missing	# Non-interviewed	# Missing
Income	42	61	77	106
Poverty Index	42	61	77	109
Total Delay	42	50	77	94
Patient Delay	42	46	77	86

Table 15: Associations between specified variables and poverty index for cases with complete information

Complete		Contingency Coefficient**	Rank of association with Poverty index
Variables	Race	0.44	1
	Stage	0.25	4
	Location	0.32	2
	Age group	0.24	5
	Grade	0.25	3
	Symptoms	0.05	8
	Histology	0.22	6
	Menopausal status	0.19	7
V2*	Smoking history	0.15	6
	Type of insurance	0.50	2
	Education	0.50	3
	Oral contraceptive use	0.26	5
	Usual source of care	0.35	4
	Occupation Class	0.54	1

* Variables missing only because absence of interview

* * A measurement of association for contingency tables (Cohen, 1960)

Table 16: Log-linear models for subset of variables correlated with poverty index

		Chi-Square	Df	P-value
Complete	Race	55.31	3	<.0001
Variables	Location | Race	28.73	6	<.0001
	Grade | Race, Location	10.05	6	0.12
	Stage | Race, Location	9.99	8	0.27
	Age group | Race, Location	12.82	6	0.046
	Histology | Race, Location, Age group	2.02	3	0.57
	Menopausal status | Race, Location, Age group	0.47	2	0.79
	Symptoms | Race, Location, Age group	0		
V2*				
	Occupation Class| Race, Location, Age group	26.78	11	0.005
	Insurance| Race, Location, Age group, Occupation class	2.53	6	0.87
	Education | Race, Location, Age group, Occupation class	3.53	6	0.74
	Usual Source of Care | Race, Location, Age group, Occupation class	2.35	6	0.88
	Oral Contraceptive use| Race, Location, Age group, Occupation class	1.41	3	0.70
	Smoking history | Race, Location, Age group, Occupation class	3.39	6	0.76

* Variables missing only because absence of interview

Table 17: The cross classification of the poverty index by occupational class according to whether information is present or missing

Poverty index	Occupation class		
	Present	Missing	Total
Present	317(65%)	3(1%)	320(65%)
Missing	51(10%)	119 (24%)	170(35%)
Total	368(75%)	12(25%)	490(100%)

Table 18 : The frequency distribution of poverty index in original data and imputed data

	Original			Imputed			
	Total (N=320)	White (N= 232)	Black (N=88)		Total (N=490)	White (N=341)	Black (N=149)
Poverty index	Frequency (%)	Frequency (%)	Frequency (%)	Poverty index	Frequency (%)	Frequency (%)	Frequency (%)
0-125	32.5	19.83	65.91	0-125	37.35	23.75	68.46
126-200	10.31	9.48	12.50	126-200	10.41	10.26	10.74
201-300	13.75	13.79	13.64	201-300	14.29	14.08	14.77
>300	43.44	56.90	7.95	>300	37.96	51.91	6.04

Table 19: Hazard ratio (HR) for overall mortality and its 95% confidence interval (CI) for poverty index in original data and imputed data

	Original		Imputed	
Poverty index	HR	95% CI	HR	95% CI
0-125*	2.41	(1.33, 4.39)	3.79	(2.38, 6.04)
126-200*	1.57	(0.75, 3.29)	2.04	(1.15, 3.63)
201-300*	2.28	(1.20, 4.34)	2.86	(1.73, 4.74)

* Reference group is the group with poverty index greater than 300

Table 20: Comparison of estimated hazard ratios for overall mortality from the paper by Hill et al (1996) with those based on the updated original dataset and the multiply imputed dataset adjusting for selected covariates using Cox proportional hazards models

Variables in the Model**			
Covariate:	Source	Hazard Ratio for Race	95% Confidence Interval
Race (only)	Table 4, Hill et al (1996)*	4.0	(2.8, 5.6)
	Updated original dataset	3.3	(2.5, 4.3)
	Multiply imputed dataset	3.3	(2.5, 4.3)
Stage	Table 4, Hill et al (1996)*	2.8	(1.9, 4.0)
	Updated original dataset	2.5	(1.9, 3.4)
	Multiply imputed dataset	2.5	(1.9, 3.4)
Tumor characteristics	Table 4, Hill et al (1996)*	3.1	(2.2, 4.4)
	Updated original dataset	2.6	(2.0, 3.5)
	Multiply imputed dataset	2.6	(2.0, 3.5)
Treatment	Table 4, Hill et al (1996)*	3.1	(2.1, 4.5)
	Updated original dataset	3.0	(2.0, 3.7)
	Multiply imputed dataset	2.8	(2.1, 3.6)
Sociodemographic	Table 4, Hill et al (1996)*	3.3	(2.2, 4.9)
	Updated original dataset	1.8	(1.1, 3.0)
	Multiply imputed dataset	1.8	(1.4, 2.7)
Hormonal and reproductive	Table 4, Hill et al (1996)*	3.4	(2.3, 5.0)
	Updated original dataset	3.0	(2.0, 4.1)
	Multiply imputed dataset	3.1	(2.4, 4.1)

* Source: Hill HA et al (1996) ** All the models include age and location.

Table 20: Comparison of estimated hazard ratios for overall mortality from the paper by Hill et al (1996) with those based on the updated original dataset and the multiply imputed dataset adjusting for selected covariates using Cox proportional hazards models (continued)

Comorbidities and health behavior	Table 4, Hill et al (1996)*	3.3	(2.3, 4.9)
	Updated original dataset	2.6	(1.6, 4.0)
	Multiply imputed dataset	2.5	(2.0, 3.7)
Stage, tumor characteristics	Table 4, Hill et al (1996)*	2.5	(1.7, 3.6)
	Updated original dataset	2.3	(1.7, 3.1)
	Multiply imputed dataset	2.3	(1.7, 3.1)
Stage, tumor characteristics, treatment	Table 4, Hill et al (1996)*	2.1	(1.4, 3.2)
	Updated original dataset	2.2	(1.5, 3.0)
	Multiply imputed dataset	2.1	(1.5, 2.8)
Stage, tumor characteristics, treatment, hormonal and reproductive	Table 4, Hill et al (1996)*	1.9	(1.2, 3.0)
	Updated original dataset	2.3	(1.5, 3.4)
	Multiply imputed dataset	2.1	(1.5, 2.8)
Stage, tumor characteristics, treatment, hormonal and reproductive, comorbidities and health behavior	Table 4, Hill et al (1996)*	1.8	(1.1, 2.8)
	Updated original dataset	2.1	(1.3, 3.5)
	Multiply imputed dataset	1.7	(1.2, 2.5)
Stage, tumor characteristics, treatment, hormonal and reproductive, Comorbidities and health behavior, sociodemographic	Table 4, Hill et al (1996)*	1.6	(1.0, 2.6)
	Updated original dataset	1.3	(0.7, 2.6)
	Multiply imputed dataset	1.2	(0.9, 1.9)

* Source: Hill HA et al (1996) ** All the models include age and location.

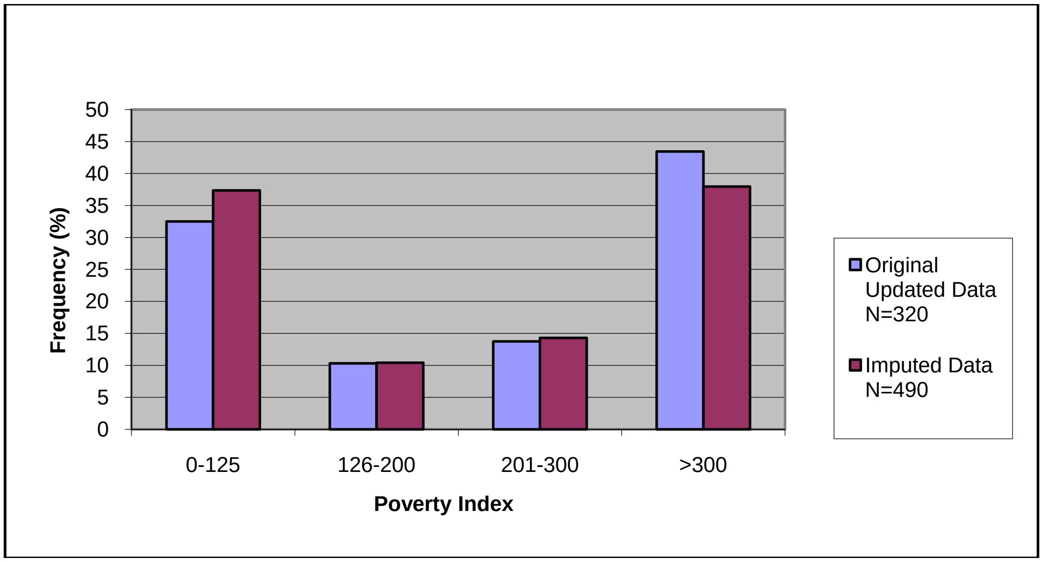

Figure 1: Distribution of poverty index in original updated and imputed dataset

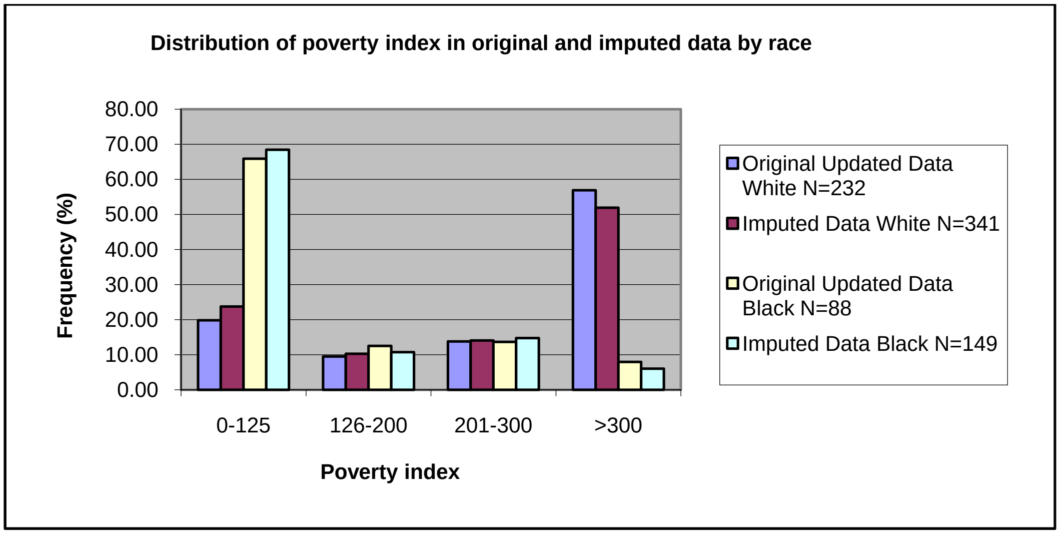

Figure 2: Distribution of poverty index in original updated and imputed dataset by race

APPENDIX A

LIST OF INFORMATION FOR ALL VARIABLES IN THE ANALYSES

Study Variable	Data Source	Categories	Original Code	Code used
Race	Medical record*	Black	1	1
		White	0	0
Location	Study Design	Atlanta	1	1
		New Orleans	2	2
		San Fran/Oak	3	3
Age Group	Study Design	20-49	1	1
		50-64	2	2
		65-79	3	3
Stage	Medical record*	Unknown	V	
		1	1	1
		2	2	2
		3	3	3
		4	4	4
Interviewed	Study Design	No	0	0
		Yes	1	1
Grade	Pathology Review	Unknown	V	
		1	1	1
		2	2	2
		3	3	3
Comorbidity	Medical record*	Unknown	V	.
		No	0	0
		Yes	1	1
Symptoms	Medical record*	Unknown	V	
		No	0	0
		Yes	1	1

* Include medical records abstract and physician records

List of information for all variables in the analyses (continue)

Study Variable	Data Source	Categories	Original Code	Code used
Patient Delay	Interview	No	0	0
		0-<1m	1	1
		1-<3m	2	2
		3-<6m	3	3
		>=6m	4	4
		Unknown	.	.
Total Delay	Interview + Medical record**	0-<1m	1	1
		1-<3m	2	2
		3-<6m	3	3
		>=6m	4	4
		Unknown	.	.
Smoking History	Interview	Never	0	0
		Former	1	1
		Current	2	2
		Unknown	.	.
Income	Interview	<10K	1	1
		10-<20K	2	2
		20-<35K	3	3
		35K	4	4
		Unknown	.	.
Type of Insurance	Interview	None	0	0
		Public	1	1
		Any	2	2
		Unknown	.	.
Usual Source of Care	Interview	None	0	0
		Public	1	1
		Private	2	2
		Unknown	.	.
Poverty Index	Interview	0-125	1	1
		126-200	2	2
		201-300	3	3
		>301-400	4	4
		>400	5	4
		Unknown	.	.
Occupation Class	Interview	Homemaker	0	5
		Mgt/Prof	1	1
		Tech/sales	2	2
		Skilled	3	3
		Unskilled	4	4
		Unknown	.	.

** Total delay was calculated by adding system delay, which was derived from medical records, and patient delay, which was derived from personal interviews.

List of information for all variables in the analyses (continue)

Study Variable	Data Source	Categories	Original Code	Code used
Education	Interview	High School Grad	2	2
		>High school	3	3
		Unknown	.	.
Marital Status	Interview	Partnered	1	1
		Widowed	2	2
		Div/Sep	3	3
		Never married	6	4
		Unknown	.	.
Body Mass Index Quartile	Medical record*	Low normal	1	1
		High normal	2	2
		Overweight	3	3
		Very overweight	4	4
		Unknown	.	.
Histology	Pathology review	Others	0	0
		Adenosquamous	1	1
Therapy	Medical record*	No Surgery	NoSurgery	1
		Surgery only	SurgOnly	2
		Surgery plus	SurgPlusChemo[+] SurgPlusHormn[++] SurgPlusCandH[+++] SurgPlusRT[**] Surg_RT_CorH[***]	3
		Unknown	TRTUnkn	.
Oral Contraceptive Use	Interview	Not used	0	0
		Used	1	1
		Unknown	.	.
Menopausal status	Medical record*	Premenopausal	0	0
		Postmenopausal	1	1
		Unknown	.	.

* Include medical records abstract and physician records

+Surgery plus chemotherapy; ++Surgery plus hormonotherapy; +++ Surgery plus chemotherapy and hormonotherapy

Surgery plus radiotherapy; * Surgery plus radiotherapy and either chemotherapy or hormonotherapy

.

APPENDIX B

GENERATION OF THE IMPUTATION CODE

For total delay:

Step 1: Association between specified variables and total delay for cases with complete information

Complete		Contingency Coefficient
Variables	Race	0.03
	Stage	0.15
	Location	0.14
	Age group	0.31
	Grade	0.08
	Symptoms	0.16
	Histology	0.06
	Menopausal status	0.25
V2*	Smoking history	0.16
	Type of insurance	0.17
	Education	0.17
	Oral Contraceptive use	0.17
	Usual Source of Care	0.19
	Occupation Class	0.15

* Variables missing only because absence of interview

Step 2: Log-linear models for subset of variables correlated with total delay

Complete		Chi-Square	Df	p-value
Variables	Age group	30.76	6	<.0001
	Menopausal status | Age group	1.49	3	0.68
	Symptoms | Age group	1.38	2	0.50
	Stage| Age group	9.38	9	0.40
	Location | Age group	6.52	6	0.37
	Grade | Age group	3.11	6	0.79
	Histology | Age group	0.97	3	0.81
	Race | Age group	0.33	3	0.95
V2*	Usual Source of Care | Age group	7.5	6	0.28
	Oral Contraceptive use | Age group	0.8	3	0.85
	Education| Age group	6.89	6	0.33
	Insurance | Age group	6.11	6	0.41
	Smoking hist | Age group	6.07	6	0.42
	Occupation Class | Age group	6.78	12	0.87

* Variables missing only because absence of interview

Use

totdlygpcode=Survival time*100+Age group*10+Location*1

to impute total delay

For patient delay:

Impute patient delay as 0 for those symptoms equal to 0

Step 1: Association between specified variables and patient delay for cases with complete information

Complete		Contingency Coefficient
Variables	Race	0.13
	Stage	0.13
	Location	0.16
	Age group	0.24
	Grade	0.21
	Symptoms	0.71
	Histology	0.10
	Menopausal status	0.19

Step 2: Log-linear models for subset of variables correlated with patient delay

		Chi-Square	Df	p-value
Complete	Age group	20.09	8	0.01
Variables	Grade | Age group	7.11	8	0.53
	Menopausal | Age group	2.56	4	0.63
	Location | Age group	7.67	8	0.47
	Race | Age group	2.61	4	0.62
	Stage | Age group	6.81	11	0.81
	Histology | Age group	1.94	4	0.75

Use

ptdlygpcode= Survival time*100+ Age group*10+ Location*1

to impute patient delay.

For smoking history:

Step 1: Association between specified variables and smoking history for cases with complete information

Complete		Contingency Coefficient
Variables	Race	0.11
	Stage	0.13
	Location	0.10
	Age group	0.25
	Grade	0.15
	Symptoms	0.04
	Histology	0.05
	Menopausal status	0.11

Step 2: Log-linear models for subset of variables correlated with smoking history

		Chi-Square	Df	p-value
Complete	Age group	20.82	4	0.0003
Variables	Grade | Age group	10.06	4	0.0395
	Stage | Age group, Grade	3.02	6	0.8067
	Menopausal status | Age group, Grade	2.09	2	0.3512
	Race | Age group, Grade	2.77	2	0.2504
	Location | Age group, Grade	1.17	4	0.8824
	Histology | Age group, Grade	2.22	2	0.3294
	Symptoms | Age group, Grade	2.13	2	0.3447

Use

smkhxcode=Survival time*1000+Age group*100+Grade*10+Location*1

to impute smoking history

For type of insurance:

Step 1: Association between specified variables and type of insurance for cases with complete information

Complete		Contingency Coefficient
Variables	Race	0.38
	Stage	0.14
	Location	0.21
	Age group	0.32
	Grade	0.15
	Symptoms	0.02
	Histology	0.17
	Menopausal status	0.19

Step 2: Log-linear models for subset of variables correlated with type of insurance

		Chi-Square	Df	p-value
Complete	Race	55.01	2	<.0001
Variables	Age group| Race	26.61	4	<.0001
	Location| Race, Age group	9.67	4	0.046
	Menopausal status | Race, Age group, Location	2.6	1	0.11
	Histology| Race, Age group, Location	2.91	2	0.23
	Grade | Race, Age group, Location	0.92	4	0.92
	Stage | Race, Age group, Location	5.17	6	0.52
	Symptoms | Race, Age group, Location	1.09	2	0.58

Use

insgpcode=Survival time*1000+ Race*100+ Age group*10+ Location*1

to impute insurance.

For education:

Step 1: Association between specified variables and education for cases with complete information

Complete		Contingency Coefficient
Variables	Race	0.35
	Stage	0.14
	Location	0.23
	Age group	0.23
	Grade	0.17
	Symptoms	0.05
	Histology	0.16
	Menopausal status	0.23

Step 2: Log-linear models for subset of variables correlated with education

		Chi-Square	Df	p-value
Complete	Race	45.39	2	<.0001
Variables	Menopausal status | Race	18.96	2	<.0001
	Age group| Race, Menopausal status	1.52	4	0.82
	Location| Race, Menopausal status	14.24	4	0.0066
	Grade| Race, Menopausal status, Location	2.94	4	0.57
	Histology| Race, Menopausal status, Location	0.91	2	0.63
	Stage| Race, Menopausal status, Location	1.67	6	0.95
	Symptoms| Race, Menopausal status, Location	0.67	2	0.72

Use

educcode= Survival time*1000+ Race*100+ Age group*10+ Location*1

to impute education.

For oral contraceptive use:

Step 1: Association between specified variables and oral contraceptive use for cases with complete information

Complete		Contingency Coefficient
Variables	Race	0.14
	Stage	0.14
	Location	0.15
	Age group	0.48
	Grade	0.20
	Symptoms	0.03
	Histology	0.01
	Menopausal status	0.48

Step 2: Log-linear models for subset of variables correlated with oral contraceptive use

		Chi-Square	Df	p-value
Complete	Menopausal status	82.15	1	<.0001
Variables	Age group | Menopausal status	16.93	2	0.0002
	Grade | Menopausal status, Age group	8.24	2	0.016
	Location| Age group, Menopausal status, Grade	1.44	2	0.49
	Race | Age group, Menopausal status, Grade	3.19	1	0.074
	Stage | Age group, Menopausal status, Grade	1.96	3	0.58
	Symptoms | Age group, Menopausal status, Grade	0.87	1	0.35
	Histology | Age group, Menopausal status, Grade	0.01	1	0.94

Use

ocpcode=Survival time*1000+ Age group*100+ Menopausal status*10+ Grade*1

to impute oral contraceptive use

For usual source of care:

Step 1: Association between specified variables and usual source of care for cases with complete information

Complete		Contingency Coefficient
Variables	Race	0.38
	Stage	0.22
	Location	0.22
	Age group	0.18
	Grade	0.10
	Symptoms	0.04
	Histology	0.14
	Menopausal Status	0.14

Step 2: Log-linear models for subset of variables correlated with oral contraceptive use

		Chi-Square	Df	p-value
Complete	Race	42.51	2	<.0001
Variables	Stage | Race	16.21	6	0.013
	Location | Race, Stage	13.53	4	0.009
	Age group | Race, Stage, Location	10.25	4	0.036
	Histology | Race, Stage, Location, Age group	0.16	2	0.92
	Menopausal status | Race, Stage, Location, Age group	0.67	2	0.71
	Grade | Race, Stage, Location, Age group	5.87	4	0.21
	Symptoms | Race, Stage, Location, Age group	1.63	2	0.44

Use

ucgpcode=Survivaltime*10000+Race*1000+Stage*100+Location*10+Agegroup*1

to impute usual source of care

For occupation class:

Step 1: Association between specified variables and occupation class for cases with complete information

Complete		Contingency Coefficient
Variables	Race	0.40
	Stage	0.27
	Location	0.25
	Age group	0.25
	Grade	0.25
	Symptoms	0.07
	Histology	0.20
	Menopausal Status	0.23

Step 2: Log-linear models for subset of variables correlated with occupation class

		Chi-Square	Df	p-value
Complete	Race	61.08	4	<.0001
Variables	Stage | Race	11.97	11	0.37
	Location | Race	22.28	8	0.0044
	Age group | Race, Location	12.17	8	0.14
	Grade | Race, Location	6.54	8	0.59
	Menopausal status | Race, Location	4.24	4	0.37
	Histology | Race, Location	8.26	4	0.082
	Symptoms | Race, Location	2.99	4	0.56

Use

occupcode= Survival time*1000+ Race*100+ Location*10+ Age group*1

to impute occupation class.

For comorbidity:

Step 1: Association between specified variables and comorbidity for cases with complete information

Complete		Contingency Coefficient
Variables	Race	0.20
	Stage	0.09
	Location	0.07
	Age group	0.22
	Grade	0.09
	Symptoms	0.03
	Histology	0.08
	Menopausal Status	0.22

Step 2: Log-linear models for subset of variables correlated with comorbidity

		Chi-Square	Df	p-value
Complete	Menopausal status	20.38	1	<.0001
Variables	Age group | Menopausal status	0.52	2	0.77
	Race | Menopausal status	17.11	1	<.0001
	Stage | Menopausal status, Race	2.15	3	0.54
	Grade | Menopausal status, Race	1.22	2	0.54
	Histology | Menopausal status, Race	7.32	1	0.0068
	Location | Menopausal status, Race, Histology	2.52	2	0.28
	Symptoms | Menopausal status, Race, Histology	0	1	0.97

Use

comorbscode= Survival time*1000+ Age group*100+ Race*10+Histology*1

to impute comorbidity.

For BMI:

Step 1: Association between specified variables and BMI for cases with complete information

Complete		Contingency Coefficient
Variables	Race	0.38
	Stage	0.19
	Location	0.22
	Age group	0.14
	Grade	0.15
	Symptoms	0.06
	Histology	0.15
	Menopausal Status	0.12

Step 2: Log-linear models for subset of variables correlated with BMI

		Chi-Square	Df	p-value
Complete	Race	57.87	3	<.0001
Variables	Location | Race	13.03	6	0.0426
	Stage | Race, Location	4.18	9	0.8993
	Grade| Race, Location	5.09	6	0.5329
	Histology | Race, Location	2.14	3	0.5431
	Age group | Race, Location	6.04	6	0.4182
	Menopausal status | Race, Location	4.92	3	0.1775
	Symptoms | Race, Location	1.08	3	0.7828

Use

bmiqcode= Survival time*1000+ Race*100+ Location*10+ Age group*1

to impute BMI

For treatment:

Step 1: Association between specified variables and treatment for cases with complete information

Complete		Contingency Coefficient
Variables	Race	0.31
	Stage	0.43
	Location	0.20
	Age group	0.13
	Grade	0.41
	Symptoms	0.09
	Histology	0.20
	Menopausal Status	0.12

Step 2: Log-linear models for subset of variables correlated with treatment

		Chi-Square	Df	p-value
Complete	Stage	58.75	6	<.0001
Variables	Grade| Stage	27.94	6	<.0001
	Race| Stage, Grade	14.55	2	0.0007
	Location | Stage, Grade, Race	12.65	4	0.013
	Histology | Stage, Grade, Race, Location	1.2	2	0.55
	Age group | Stage, Grade, Race, Location	1.16	4	0.88
	Menopausal status | Stage, Grade, Race, Location	1.29	2	0.52
	Symptoms | Stage, Grade, Race, Location	0.18	1	0.67

Use

therapycode= Survival time*10000+ Stage*1000+ Grade*100+ Race*10+ Location*1

to impute treatment

APPENDIX C

SAS CODE FOR THE GENERATION OF IMPUTATION CODE AND R CODE FOR MULTIPLE IMPUTATION

C.1 SAS CODE FOR THE GENERATION IMPUTATION CODE

```
*For generation imputation code;
*First, look at the missing distribution;
proc sort data=sasuser.labeled_dum;
by race;
run;
proc freq data=sasuser.labeled_dum;
by race;
table
inques*ptdlygpM
inques*totdlygpM
inques*smkhxM
inques*incgpM
inques*insgpM
inques*ucgpM
inques*povgpM
inques*occupM
inques*educM
inques*ocpM
;
run;

* Impute almost complete variables;

*Association between menostat and agegp;
proc freq data=sasuser.impute;
tables menostat*agegp/cmh;
run;
proc freq data=sasuser.impute;
tables menostatM*agegp;
```

```
run;

Data sasuser.baseline;
set sasuser.labeledm;
if survtime=<12 then st2=1;
else if survtime=<36 & survtime >12 then st2=2;
else if survtime=<60 & survtime >36 then st2=3;
else st2=4;
if stage=. then stageM=1; else stageM=0;
if grade=. then gradeM=1; else gradeM=0;
proc freq;
tables stage*st2 stageM*st2 grade*st2 gradeM*st2/chisq cmh;
run;
data t;
set sasuser.labeledm;
if symptoms=0 then symptomsM=0;
else if symptoms=1 then symptomsM=0;
else symptomsM=1;
run;
proc freq;
tables symptomsM*ptdlygp;
run;
proc freq data=sasuser.labeled_dum;
table menostatM*agegp;
run;

* Merge variables to avoid sparse data;
data sasuser.impute;
set sasuser.impute;
if povgp=1 then povgp=1;
else if povgp=2 then povgp=2;
else if povgp=3 then povgp=3;
else if povgp=4 then povgp=4;
else if povgp=5 then povgp=4;
else povgp=.;
if stage=. then stage=1;
if grade=.& st2=1 then grade=3;
else if grade=. & st2=3 then grade=3;
else if grade=. & st2=4 then grade=1;
if symptoms=. then symptoms=1;
if menostat=.& agegp=1 then menostat=1;
else if menostat=. & agegp=2 then menostat=2;
run;
proc freq;
table povgp*race;
run;

*1. Povery index;
* pairwise correlation with basic variales by contingency table
analysis for poverty index;
proc freq;
table povgp*race/chisq cmh;
run;
proc freq;
table povgp*stage/chisq cmh;
run;
proc freq;
```

```
table povgp*locn/chisq cmh;
run;
proc freq;
table povgp*agegp/chisq cmh;
run;
proc freq;
table povgp*grade/chisq cmh;
run;
proc freq;
table povgp*symptoms/chisq cmh;
run;
proc freq;
table povgp*Histcat/chisq cmh;
run;
proc freq;
table povgp*MENOSTAT/chisq cmh;
run;
* Pairwise with V2 for poverty index;
proc freq;
table povgp*smkhx/chisq cmh;
run;
proc freq;
table povgp*insgp/chisq cmh;
run;
proc freq;
table povgp*educ/chisq cmh;
run;
proc freq;
table povgp*OCP/chisq cmh;
run;
proc freq;
table povgp*ucgp/chisq cmh;
run;
proc freq;
table povgp*occup/chisq cmh;
run;

* loglinear models with baseline variables for poverty index;
proc catmod;
model povgp*race=_response_;
loglin povgp|race;
run;
proc catmod;
model povgp*race*locn=_response_;
loglin povgp|race|locn @2;
run;
proc catmod;
model povgp*race*locn*grade=_response_ / noparm;
loglin povgp|race|locn|grade @2;
run;
proc catmod;
model povgp*race*locn*stage=_response_ / noparm;
loglin povgp|race|locn|stage @2;
run;
proc catmod;
model povgp*race*locn*agegp=_response_ / noparm;
loglin povgp|race|locn|agegp @2;
```

```
run;
proc catmod;
model povgp*race*locn*agegp*histcat=_response_ / noparm;
loglin povgp|race|locn|agegp|histcat @2;
run;
proc catmod;
model povgp*race*locn*agegp*menostat=_response_ / noparm;
loglin povgp|race|locn|agegp|menostat @2;
run;
proc catmod;
model povgp*race*locn*agegp*symtoms=_response_ / noparm;
loglin povgp|race|locn|agegp|symtoms @2;
run;

* loglinear models with V2 variables for poverty index;
proc catmod;
model povgp*locn*race*agegp*occup=_response_ /noparm;
loglin povgp|agegp|race|locn|occup @2;
run;
proc catmod;
model povgp*locn*race*agegp*occup*insgp=_response_ /noparm;
loglin povgp|locn|race|agegp|occup|insgp @2;
run;
proc catmod;
model povgp*locn*race*agegp*occup*educ=_response_ /noparm;
loglin povgp|locn|race|agegp|occup|educ @2;
run;
proc catmod;
model povgp*locn*race*agegp*occup*ucgp=_response_ /noparm;
loglin povgp|locn|race|agegp|occup|ucgp @2;
run;
proc catmod;
model povgp*locn*race*agegp*occup*ocp=_response_ /noparm;
loglin povgp|locn|race|agegp|occup|ocp @2;
run;
proc catmod;
model povgp*locn*race*agegp*occup*smkhx=_response_ /noparm;
loglin povgp|locn|race|agegp|occup|smkhx @2;
run;

*Prepare for imputation;
proc freq data=sasuser.impute;
table povgpM*occupM;
run;
proc corr;
var occup educ;
run;
proc sort data=sasuser.impute;
by locn race;
proc freq;
tables occup*povgp;
by race;
run;
proc freq;
tables occup*povgp;
run;
proc freq;
```

```
tables occup*povgp;
by locn;
run;
proc freq;
tables occup*povgp;
by locn race;
run;

data sasuser.impute;
set sasuser.impute;
id=_N_;
povgpcode=st2*10000+race*1000+locn*100+agegp*10+occup*1;
povgpcode2=st2*1000+race*100+locn*10+agegp*1;
run;

*2. Total delay;
* pairwise correlation with basic variales by contingency table
analysis for total delay;
proc freq;
table totdlygp*race/chisq cmh;
run;
proc freq;
table totdlygp*stage/chisq cmh;
run;
proc freq;
table totdlygp*locn/chisq cmh;
run;
proc freq;
table totdlygp*agegp/chisq cmh;
run;
proc freq;
table totdlygp*grade/chisq cmh;
run;
proc freq;
table totdlygp*symptoms/chisq cmh;
run;
proc freq;
table totdlygp*Histcat/chisq cmh;
run;
proc freq;
table totdlygp*MENOSTAT/chisq cmh;
run;
* Pairwise with V2 for total delay;
proc freq;
table totdlygp*smkhx/chisq cmh;
run;
proc freq;
table totdlygp*insgp/chisq cmh;
run;
proc freq;
table totdlygp*educ/chisq cmh;
run;
proc freq;
table totdlygp*OCP/chisq cmh;
run;
proc freq;
table totdlygp*ucgp/chisq cmh;
```

```
run;
proc freq;
table totdlygp*occup/chisq cmh;
run;
* loglinear models with baseline variables for total delay;
proc catmod;
model totdlygp*agegp=_response_;
loglin totdlygp|agegp;
run;
proc catmod;
model totdlygp*agegp*menostat=_response_ / noparm;
loglin totdlygp|agegp|menostat @2;
run;
proc catmod;
model totdlygp*agegp*symptoms=_response_ / noparm;
loglin totdlygp|symptoms|agegp @2;
run;
proc catmod;
model totdlygp*agegp*stage=_response_ / noparm;
loglin totdlygp|agegp|stage @2;
run;
proc catmod;
model totdlygp*agegp*locn=_response_ / noparm;
loglin totdlygp|agegp|locn @2;
run;
proc catmod;
model totdlygp*agegp*grade=_response_ / noparm;
loglin totdlygp|agegp|grade @2;
run;
proc catmod;
model totdlygp*agegp*histcat=_response_ / noparm;
loglin totdlygp|agegp|histcat @2;
run;
proc catmod;
model totdlygp*agegp*race=_response_ / noparm;
loglin totdlygp|agegp|race @2;
run;
* loglinear models with V2 variables for total delay;
proc catmod;
model totdlygp*agegp*ucgp=_response_ /noparm;
loglin totdlygp|agegp|ucgp @2;
run;
proc catmod;
model totdlygp*agegp*ocp=_response_ /noparm;
loglin totdlygp|agegp|ocp @2;
run;
proc catmod;
model totdlygp*agegp*educ=_response_ /noparm;
loglin totdlygp|agegp|educ @2;
run;
proc catmod;
model totdlygp*agegp*insgp=_response_ /noparm;
loglin totdlygp|agegp|insgp @2;
run;
proc catmod;
model totdlygp*agegp*smkhx=_response_ /noparm;
loglin totdlygp|agegp|smkhx @2;
```

```
run;
proc catmod;
model totdlygp*agegp*occup=_response_ /noparm;
loglin totdlygp|agegp|occup @2;
run;

*Prepare for imputation;
data sasuser.impute;
set sasuser.impute;
totdlygpcode=st2*100+Agegp*10+locn*1;
run;

* 3. Smoking history;
*  pairwise  correlation  with  basic  variales  by  contingency  table
analysis for smoking history;
proc freq;
table smkhx*race/chisq cmh;
run;
proc freq;
table smkhx*stage/chisq cmh;
run;
proc freq;
table smkhx*locn/chisq cmh;
run;
proc freq;
table smkhx*agegp/chisq cmh;
run;
proc freq;
table smkhx*grade/chisq cmh;
run;
proc freq;
table smkhx*symptoms/chisq cmh;
run;
proc freq;
table smkhx*Histcat/chisq cmh;
run;
proc freq;
table smkhx*MENOSTAT/chisq cmh;
run;

* loglinear models with baseline variables for smoking history;
proc catmod;
model smkhx*agegp=_response_;
loglin smkhx|agegp;
run;
proc catmod;
model smkhx*agegp*grade=_response_ / noparm;
loglin smkhx|agegp|grade @2;
run;
proc catmod;
model smkhx*agegp*grade*stage=_response_ / noparm;
loglin smkhx|agegp|grade|stage @2;
run;
proc catmod;
model smkhx*agegp*grade*menostat=_response_ / noparm;
loglin smkhx|agegp|grade|menostat @2;
run;
```

```
proc catmod;
model smkhx*agegp*grade*race=_response_ / noparm;
loglin smkhx|agegp|grade|race @2;
run;
proc catmod;
model smkhx*agegp*grade*locn=_response_ / noparm;
loglin smkhx|agegp|grade|locn @2;
run;
proc catmod;
model smkhx*agegp*grade*histcat=_response_ / noparm;
loglin smkhx|agegp|grade|histcat @2;
run;
proc catmod;
model smkhx*agegp*grade*symptoms=_response_ / noparm;
loglin smkhx|agegp|grade|symptoms @2;
run;

*Prepare for imputation;
data sasuser.impute;
set sasuser.impute;
smkhxcode=st2*1000+Agegp*100+grade*10+locn*1;
run;

*4. Insurance;
* pairwise correlation with basic variales by contingency table
analysis for insurance;
proc freq;
table insgp*race/chisq cmh;
run;
proc freq;
table insgp*stage/chisq cmh;
run;
proc freq;
table insgp*locn/chisq cmh;
run;
proc freq;
table insgp*agegp/chisq cmh;
run;
proc freq;
table insgp*grade/chisq cmh;
run;
proc freq;
table insgp*symptoms/chisq cmh;
run;
proc freq;
table insgp*Histcat/chisq cmh;
run;
proc freq;
table insgp*MENOSTAT/chisq cmh;
run;

* loglinear models with baseline variables for insurance;
proc catmod;
model insgp*race=_response_;
loglin insgp|race;
run;
proc catmod;
```

```
model insgp*race*agegp=_response_ / noparm;
loglin insgp|race|agegp @2;
run;
proc catmod;
model insgp*race*agegp*locn=_response_ / noparm;
loglin insgp|race|agegp|locn @2;
run;
proc catmod;
model insgp*race*agegp*locn*menostat=_response_ / noparm;
loglin insgp|race|agegp|locn|menostat @2;
run;
proc catmod;
model insgp*race*agegp*locn*histcat=_response_ / noparm;
loglin insgp|race|agegp|locn|histcat @2;
run;
proc catmod;
model insgp*race*agegp*locn*grade=_response_ / noparm;
loglin insgp|race|agegp|locn|grade @2;
run;
proc catmod;
model insgp*race*agegp*locn*stage=_response_ / noparm;
loglin insgp|race|agegp|locn|stage @2;
run;
proc catmod;
model insgp*race*locn*agegp*symptoms=_response_ / noparm;
loglin insgp|race|locn|agegp|symptoms @2;
run;

*Prepare for imputation;
data sasuser.impute;
set sasuser.impute;
insgpcode=st2*1000+race*100+agegp*10+locn*1;
run;

*5. Education;
* pairwise correlation with basic variales by contingency table
analysis for Education;
proc freq;
table educ*race/chisq cmh;
run;
proc freq;
table educ*stage/chisq cmh;
run;
proc freq;
table educ*locn/chisq cmh;
run;
proc freq;
table educ*agegp/chisq cmh;
run;
proc freq;
table educ*grade/chisq cmh;
run;
proc freq;
table educ*symptoms/chisq cmh;
run;
proc freq;
table educ*Histcat/chisq cmh;
```

```
run;
proc freq;
table educ*MENOSTAT/chisq cmh;
run;

* loglinear models with baseline variables for education;
proc catmod;
model educ*race=_response_;
loglin educ|race;
run;
proc catmod;
model educ*race*menostat=_response_ / noparm;
loglin educ|race|menostat @2;
run;
proc catmod;
model educ*race*menostat*agegp=_response_ / noparm;
loglin educ|race|menostat|agegp @2;
run;
proc catmod;
model educ*race*menostat*locn=_response_ / noparm;
loglin educ|race|menostat|locn @2;
run;
proc catmod;
model educ*race*menostat*locn*grade=_response_ / noparm;
loglin educ|race|menostat|locn|grade @2;
run;
proc catmod;
model educ*race*menostat*locn*histcat=_response_ / noparm;
loglin educ|race|menostat|locn|histcat @2;
run;
proc catmod;
model educ*race*menostat*locn*stage=_response_ / noparm;
loglin educ|race|menostat|locn|stage @2;
run;
proc catmod;
model educ*race*menostat*locn*symptoms=_response_ / noparm;
loglin educ|race|menostat|locn|symptoms @2;
run;

*Prepare for imputation;
data sasuser.impute;
set sasuser.impute;
educcode=st2*1000+race*100+agegp*10+locn*1;
run;

*6. Oral contraceptive use;
* pairwise correlation with basic variales by contingency table
analysis for OCP;
proc freq;
table ocp*race/chisq cmh;
run;
proc freq;
table ocp*stage/chisq cmh;
run;
proc freq;
table ocp*locn/chisq cmh;
run;
```

```
proc freq;
table ocp*agegp/chisq cmh;
run;
proc freq;
table ocp*grade/chisq cmh;
run;
proc freq;
table ocp*symptoms/chisq cmh;
run;
proc freq;
table ocp*Histcat/chisq cmh;
run;
proc freq;
table ocp*MENOSTAT/chisq cmh;
run;

* loglinear models with baseline variables for ocp;
proc catmod;
model ocp*menostat=_response_;
loglin ocp|menostat;
run;
proc catmod;
model ocp*menostat*agegp=_response_ / noparm;
loglin ocp|menostat|agegp @2;
run;
proc catmod;
model ocp*agegp*menostat*grade=_response_ / noparm;
loglin ocp|agegp|menostat|grade @2;
run;
proc catmod;
model ocp*agegp*menostat*grade*locn=_response_ / noparm;
loglin ocp|agegp|menostat|grade|locn @2;
run;
proc catmod;
model ocp*agegp*menostat*grade*race=_response_ / noparm;
loglin ocp|agegp|menostat|grade|race @2;
run;
proc catmod;
model ocp*agegp*menostat*grade*stage=_response_ / noparm;
loglin ocp|agegp|menostat|grade|stage @2;
run;
proc catmod;
model ocp*agegp*menostat*grade*symptoms=_response_ / noparm;
loglin ocp|agegp|menostat|grade|symptoms @2;
run;
proc catmod;
model  ocp*agegp*menostat*grade*histcat=_response_ / noparm;
loglin  ocp|agegp|menostat|grade|histcat @2;
run;

*Prepare for imputation;
data sasuser.impute;
set sasuser.impute;
ocpcode=st2*1000+agegp*100+menostat*10+grade*1;
run;

*7. Usual source of care;
```

```
* pairwise correlation with basic variales by contingency table
analysis for usual source of care;
proc freq;
table ucgp*race/chisq cmh;
run;
proc freq;
table ucgp*stage/chisq cmh;
run;
proc freq;
table ucgp*locn/chisq cmh;
run;
proc freq;
table ucgp*agegp/chisq cmh;
run;
proc freq;
table ucgp*grade/chisq cmh;
run;
proc freq;
table ucgp*symptoms/chisq cmh;
run;
proc freq;
table ucgp*Histcat/chisq cmh;
run;
proc freq;
table ucgp*MENOSTAT/chisq cmh;
run;

* loglinear models with baseline variables for usual source of
care;
proc catmod;
model ucgp*race=_response_;
loglin ucgp|race;
run;
proc catmod;
model ucgp*race*stage=_response_ / noparm;
loglin ucgp|race|stage @2;
run;
proc catmod;
model ucgp*race*stage*locn=_response_ / noparm;
loglin ucgp|race|stage|locn @2;
run;
proc catmod;
model ucgp*race*stage*locn*agegp=_response_ / noparm;
loglin ucgp|race|stage|locn|agegp @2;
run;
proc catmod;
model ucgp*race*stage*locn*agegp*histcat=_response_ / noparm;
loglin ucgp|race|stage|locn|agegp|histcat @2;
run;
proc catmod;
model ucgp*race*stage*locn*agegp*menostat=_response_ / noparm;
loglin ucgp|race|stage|locn|agegp|menostat @2;
run;
proc catmod;
model ucgp*race*stage*locn*agegp*grade=_response_ / noparm;
loglin ucgp|race|stage|locn|agegp|grade @2;
run;
```

```
proc catmod;
model ucgp*race*stage*locn*agegp*symptoms=_response_ / noparm;
loglin ucgp|race|stage|locn|agegp|symptoms @2;
run;

*Prepare for imputation;
data sasuser.impute;
set sasuser.impute;
ucgpcode=st2*10000+race*1000+stage*100+locn*10+Agegp*1;
run;

*8. Occupation class;
* pairwise correlation with basic variales by contingency table
analysis for occupation class;
proc freq;
table occup*race/chisq cmh;
run;
proc freq;
table occup*stage/chisq cmh;
run;
proc freq;
table occup*locn/chisq cmh;
run;
proc freq;
table occup*agegp/chisq cmh;
run;
proc freq;
table occup*grade/chisq cmh;
run;
proc freq;
table occup*symptoms/chisq cmh;
run;
proc freq;
table occup*Histcat/chisq cmh;
run;
proc freq;
table occup*MENOSTAT/chisq cmh;
run;

* loglinear models with baseline variables for occupaton class;
proc catmod;
model occup*race=_response_;
loglin occup|race;
run;
proc catmod;
model occup*race*stage=_response_ / noparm;
loglin occup|race|stage @2;
run;
proc catmod;
model occup*race*locn=_response_ / noparm;
loglin occup|race|locn @2;
run;
proc catmod;
model occup*race*locn*agegp=_response_ / noparm;
loglin occup|race|locn|agegp @2;
run;
proc catmod;
```

```
model occup*race*locn*grade=_response_ / noparm;
loglin occup|race|locn|grade @2;
run;
proc catmod;
model occup*race*locn*menostat=_response_ / noparm;
loglin occup|race|locn|menostat @2;
run;
proc catmod;
model occup*race*locn*histcat=_response_ / noparm;
loglin occup|race|locn|histcat @2;
run;
proc catmod;
model occup*race*locn*symptoms=_response_ / noparm;
loglin occup|race|locn|symptoms @2;
run;

*Prepare for imputation;
data sasuser.impute;
set sasuser.impute;
occupcode=st2*1000+race*100+locn*10+agegp*1;
run;

* Others:
*9. Comorbidity;
* pairwise correlation with basic variales by contingency table
analysis for comorbidity;
proc freq;
table comorbs*race/chisq cmh;
run;
proc freq;
table comorbs*stage/chisq cmh;
run;
proc freq;
table comorbs*locn/chisq cmh;
run;
proc freq;
table comorbs*agegp/chisq cmh;
run;
proc freq;
table comorbs*grade/chisq cmh;
run;
proc freq;
table comorbs*symptoms/chisq cmh;
run;
proc freq;
table comorbs*Histcat/chisq cmh;
run;
proc freq;
table comorbs*MENOSTAT/chisq cmh;
run;

* loglinear models with baseline variables for comorbidity;
proc catmod;
model comorbs*menostat=_response_;
loglin comorbs|menostat;
run;
proc catmod;
```

```
model comorbs*menostat*agegp=_response_ / noparm;
loglin comorbs|menostat|agegp @2;
run;
proc catmod;
model comorbs*menostat*race=_response_ / noparm;
loglin comorbs|menostat|race @2;
run;
proc catmod;
model comorbs*menostat*race*stage=_response_ / noparm;
loglin comorbs|menostat|race|stage @2;
run;
proc catmod;
model comorbs*menostat*race*grade=_response_ / noparm;
loglin comorbs|menostat|race|grade @2;
run;
proc catmod;
model comorbs*menostat*race*histcat=_response_ / noparm;
loglin comorbs|menostat|race|histcat @2;
run;
proc catmod;
model comorbs*menostat*race*histcat*locn=_response_ / noparm;
loglin comorbs|menostat|race|histcat|locn @2;
run;
proc catmod;
model comorbs*menostat*race*histcat*symptoms=_response_ / noparm;
loglin comorbs|menostat|race|histcat|symptoms @2;
run;

*Prepare for imputation;
data sasuser.impute;
set sasuser.impute;
comorbscode=st2*1000+agegp*100+race*10+histcat*1;
run;

*9. BMI;
* pairwise correlation with basic variales by contingency table
analysis for BMI;
proc freq;
table bmiq*race/chisq cmh;
run;
proc freq;
table bmiq*stage/chisq cmh;
run;
proc freq;
table bmiq*locn/chisq cmh;
run;
proc freq;
table bmiq*agegp/chisq cmh;
run;
proc freq;
table bmiq*grade/chisq cmh;
run;
proc freq;
table bmiq*symptoms/chisq cmh;
run;
proc freq;
table bmiq*Histcat/chisq cmh;
```

```
run;
proc freq;
table bmiq*MENOSTAT/chisq cmh;
run;

* loglinear models with baseline variables for occupaton class;
proc catmod;
model bmiq*race=_response_;
loglin bmiq|race;
run;
proc catmod;
model bmiq*race*locn=_response_ / noparm;
loglin bmiq|race|locn @2;
run;
proc catmod;
model bmiq*race*locn*stage=_response_ / noparm;
loglin  bmiq|race|locn|stage @2;
run;
proc catmod;
model bmiq*race*locn*grade=_response_ / noparm;
loglin bmiq|race|locn|grade @2;
run;
proc catmod;
model bmiq*race*locn*histcat=_response_ / noparm;
loglin bmiq|race|locn|histcat @2;
run;
proc catmod;
model bmiq*race*locn*agegp=_response_ / noparm;
loglin bmiq|race|locn|agegp @2;
run;
proc catmod;
model bmiq*race*locn*menostat=_response_ / noparm;
loglin bmiq|race|locn|menostat @2;
run;
proc catmod;
model bmiq*race*locn*symptoms=_response_ / noparm;
loglin bmiq|race|locn|symptoms @2;
run;

*Prepare for imputation;
data sasuser.impute;
set sasuser.impute;
bmiqcode=st2*1000+race*100+locn*10+agegp*1;
run;

*9. Patient delay;
* pairwise correlation with basic variales by contingency table
analysis for Patient delay;
proc freq;
table ptdlygp*race/chisq cmh;
run;
proc freq;
table ptdlygp*stage/chisq cmh;
run;
proc freq;
table ptdlygp*locn/chisq cmh;
run;
```

```
proc freq;
table ptdlygp*agegp/chisq cmh;
run;
proc freq;
table ptdlygp*grade/chisq cmh;
run;
proc freq;
table ptdlygp*symptoms/chisq cmh;
run;
proc freq;
table ptdlygp*Histcat/chisq cmh;
run;
proc freq;
table ptdlygp*MENOSTAT/chisq cmh;
run;

* loglinear models with baseline variables for Patient delay;
proc catmod;
model ptdlygp*symptoms=_response_;
loglin ptdlygp|symptoms;
run;
proc freq;
tables ptdlygp*symptoms;
run;
data sasuser.impute;
set sasuser.impute;
if symptoms=0 then ptdlygp=0;
run;
proc catmod;
model ptdlygp*agegp=_response_ / noparm;
loglin ptdlygp|agegp @2;
run;
proc catmod;
model ptdlygp*agegp*grade=_response_ / noparm;
loglin  ptdlygp|agegp|grade @2;
run;
proc catmod;
model ptdlygp*agegp*menostat=_response_ / noparm;
loglin ptdlygp|agegp|menostat @2;
run;
proc catmod;
model  ptdlygp*agegp*locn=_response_ / noparm;
loglin ptdlygp|agegp|locn @2;
run;
proc catmod;
model ptdlygp*agegp*race=_response_ / noparm;
loglin ptdlygp|agegp|race @2;
run;
proc catmod;
model ptdlygp*agegp*stage=_response_ / noparm;
loglin ptdlygp|agegp|stage @2;
run;
proc catmod;
model ptdlygp*agegp*histcat=_response_ / noparm;
loglin ptdlygp|agegp|histcat @2;
run;
```

```
*Prepare for imputation;
data sasuser.impute;
set sasuser.impute;
ptdlygpcode=st2*100+agegp*10+locn*1;
run;

*10. Therapy;
* pairwise correlation with basic variales by contingency table
analysis for Therapy;
proc freq;
table therapy*race/chisq cmh;
run;
proc freq;
table therapy*stage/chisq cmh;
run;
proc freq;
table therapy*locn/chisq cmh;
run;
proc freq;
table therapy*agegp/chisq cmh;
run;
proc freq;
table therapy*grade/chisq cmh;
run;
proc freq;
table therapy*symptoms/chisq cmh;
run;
proc freq;
table therapy*Histcat/chisq cmh;
run;
proc freq;
table therapy*MENOSTAT/chisq cmh;
run;

* loglinear models with baseline variables for therapy;
proc catmod;
model therapy*stage=_response_;
loglin therapy|stage;
run;
proc catmod;
model therapy*stage*grade=_response_ / noparm;
loglin therapy|stage|grade @2;
run;
proc catmod;
model therapy*stage*grade*race=_response_ / noparm;
loglin therapy|stage|grade|race @2;
run;
proc catmod;
model therapy*stage*grade*race*locn=_response_ / noparm;
loglin  therapy|stage|grade|race|locn @2;
run;
proc catmod;
model therapy*stage*grade*race*locn*histcat=_response_ / noparm;
loglin therapy|stage|grade|race|locn|histcat @2;
run;
proc catmod;
model  therapy*stage*grade*race*locn*agegp=_response_ / noparm;
```

```
loglin therapy|stage|grade|race|locn|agegp @2;
run;
proc catmod;
model therapy*stage*grade*race*locn*menostat=_response_ / noparm;
loglin therapy|stage|grade|race|locn|menostat @2;
run;
proc catmod;
model therapy*stage*grade*race*locn*symptoms=_response_ / noparm;
loglin therapy|stage|grade|race|locn|symptoms @2;
run;


*Prepare for imputation;
data sasuser.impute;
set sasuser.impute;
therapycode=st2*10000+stage*1000+grade*100+race*10+locn*1;
run;
```

C.2 R CODE FOR MULTIPLE IMPUTAION

```
      library(gdata)
      data<-read.xls("/Users/Eva/Documents/data.xls",sheet=1,verbose=F)

      # Hot-deck imputation

      searchcloseby<-function(value,set)
            {
                  distance<-abs(value-set)

                  w<-cbind(distance,set)
                  w1<-w[order(w[,1]),1:2]

                  closevalue<-as.numeric(w1[1,2])
                  #cat(value,closevalue,"\n\n\n")
                  closevalue
                  }

      #1.Poverty index

      #first 51 subjects


      povgpimpute1<-function(dataset,dataout)
      {
            #dataset<-data
            mdvector<-
dataset$id[dataset$povgpM==1&is.na(dataset$povgpcode)==F] # 51 obs
```

```
            codeset<-dataset$povgpcode[dataset$povgpM==0] # 320 obs

            codeset<-sort(codeset) # get rid of the missing
imputecodes(3)

            num.cases<-length(mdvector) # 51

            random.numbers<-runif(num.cases,0,1)

            ipovgp<-dataset$povgp

            for (i in 1:num.cases)
            {
                  icid<-mdvector[i]
                  impcode<-dataset[icid,157]

                  check<-sum(codeset==impcode)
                  if (check==0)
                  {
                        impcode<-searchcloseby(impcode,codeset)
                        }

                        subset<-
dataset$id[is.na(dataset$povgpcode)==F&dataset$povgpcode==impcode&datas
et$povgpM==0]
                  k<-length(subset)

                  selected<-ceiling(k*random.numbers[i])
                  ipovgp[icid]<-dataset[subset[selected],45]

            }

            dataout<-data.frame(data,ipovgp)
            dataout

            }
                  set.seed(100)


      # The other 119 subjects


      povgpimpute2<-function(dataset,dateout)
      {
            mdvector2<-
dataset$id[dataset$povgpM==1&is.na(dataset$povgpcode2)==F&is.na(dataset
$povgpcode)==T] # 119 obs

            codeset2<-
dataset$povgpcode2[dataset$povgpM==0|dataset$povgpM==1&is.na(dataset$po
vgpcode)==F] # 320+51 obs

            codeset2<-sort(codeset2)

            num.cases2<-length(mdvector2)

            random.numbers2<-runif(num.cases2,0,1)
            for (i in 1:num.cases2)
            {
                  icid2<-mdvector2[i]
                  impcode2<-dataset[icid2,158]

                  impcode2<-searchcloseby(impcode2,codeset2)

                  subset2<-
dataset$id[dataset$povgpM==0&dataset$povgpcode2==impcode2|dataset$povgp
M==1&is.na(dataset$povgpcode)==F&dataset$povgpcode2==impcode2]
                  k2<-length(subset2)

                  selected2<-ceiling(k2*random.numbers2[i])
                  dataset$ipovgp[icid2]<-
dataset[subset2[selected2],170]

            }
            dataout<-dataset$ipovgp
            dataout
      }

            set.seed(100)


      # total delay:

      totdlygpimpute<-function(dataset,dataout)
      {
            #dataset<-data
            mdvector<-
dataset$id[dataset$totdlygpM==1&is.na(dataset$totdlygpcode)==F]

            codeset<-dataset$totdlygpcode[dataset$totdlygpM==0]

            codeset<-sort(codeset)

            num.cases<-length(mdvector)

            random.numbers<-runif(num.cases,0,1)

            itotdlygp<-dataset$totdlygp

            for (i in 1:num.cases)
            {
                  icid<-mdvector[i]
                  impcode<-dataset[icid,159]


                  impcode<-searchcloseby(impcode,codeset)


                        subset<-
dataset$id[is.na(dataset$totdlygpcode)==F&dataset$totdlygpcode==impcode
&dataset$totdlygpM==0]
                  k<-length(subset)
                  selected<-ceiling(k*random.numbers[i])
                  itotdlygp[icid]<-dataset[subset[selected],41]

            }

            dataout<-itotdlygp
            dataout

            }
                  set.seed(100)


      # Smoking History:

      smkhximpute<-function(dataset,dataout)
      {
            #dataset<-data

            mdvector<-
dataset$id[dataset$smkhxM==1&is.na(dataset$smkhxcode)==F]

            codeset<-dataset$smkhxcode[dataset$smkhxM==0]

            codeset<-sort(codeset)

            num.cases<-length(mdvector)

            random.numbers<-runif(num.cases,0,1)

            ismkhx<-dataset$SMKHX


            for (i in 1:num.cases)
            {
                  icid<-mdvector[i]
                  impcode<-dataset[icid,160]


                  impcode<-searchcloseby(impcode,codeset)


                        subset<-
dataset$id[is.na(dataset$smkhxcode)==F&dataset$smkhxcode==impcode&datas
et$smkhxM==0]
                  k<-length(subset)

                  selected<-ceiling(k*random.numbers[i])
                  ismkhx[icid]<-dataset[subset[selected],10]

            }

            dataout<-ismkhx
            dataout

            }
                  set.seed(100)
      # Insurance:

      insgpimpute<-function(dataset,dataout)
      {
            #dataset<-data

            mdvector<-
dataset$id[dataset$insgpM==1&is.na(dataset$insgpcode)==F]

            codeset<-dataset$insgpcode[dataset$insgpM==0]

            codeset<-sort(codeset)

            num.cases<-length(mdvector)

            random.numbers<-runif(num.cases,0,1)

            iinsgp<-dataset$insgp


            for (i in 1:num.cases)
            {
                  icid<-mdvector[i]
                  impcode<-dataset[icid,161]


                  impcode<-searchcloseby(impcode,codeset)


                        subset<-
dataset$id[is.na(dataset$insgpcode)==F&dataset$insgpcode==impcode&datas
et$insgpM==0]
                  k<-length(subset)

                  selected<-ceiling(k*random.numbers[i])
                  iinsgp[icid]<-dataset[subset[selected],43]

            }

            dataout<-iinsgp
            dataout

            }
                  set.seed(100)


      # Education:

      educimpute<-function(dataset,dataout)
      {
            #dataset<-data

            mdvector<-
dataset$id[dataset$educM==1&is.na(dataset$educcode)==F]
```

```
            codeset<-dataset$educcode[dataset$educM==0]

            codeset<-sort(codeset)

            num.cases<-length(mdvector)

            random.numbers<-runif(num.cases,0,1)

            ieduc<-dataset$educ


            for (i in 1:num.cases)
            {
                  icid<-mdvector[i]
                  impcode<-dataset[icid,162]


                  impcode<-searchcloseby(impcode,codeset)


                        subset<-
dataset$id[is.na(dataset$educcode)==F&dataset$educcode==impcode&dataset
$educM==0]
                  k<-length(subset)

                  selected<-ceiling(k*random.numbers[i])
                  ieduc[icid]<-dataset[subset[selected],46]

            }

            dataout<-ieduc
            dataout

            }
                  set.seed(100)


      #Oral contraceptive use:

      ocpimpute<-function(dataset,dataout)
      {
            #dataset<-data

            mdvector<-
dataset$id[dataset$ocpM==1&is.na(dataset$ocpcode)==F]

            codeset<-dataset$ocpcode[dataset$ocpM==0]

            codeset<-sort(codeset)

            num.cases<-length(mdvector)

            random.numbers<-runif(num.cases,0,1)

            iocp<-dataset$OCP
            for (i in 1:num.cases)
            {
                  icid<-mdvector[i]
                  impcode<-dataset[icid,163]


                  impcode<-searchcloseby(impcode,codeset)


                        subset<-
dataset$id[is.na(dataset$ocpcode)==F&dataset$ocpcode==impcode&dataset$o
cpM==0]
                  k<-length(subset)

                  selected<-ceiling(k*random.numbers[i])
                  iocp[icid]<-dataset[subset[selected],26]

            }

            dataout<-iocp
            dataout

            }
                  set.seed(100)

      # Usual source of care:

      ucgpimpute<-function(dataset,dataout)
      {
            #dataset<-data

            mdvector<-
dataset$id[dataset$ucgpM==1&is.na(dataset$ucgpcode)==F]

            codeset<-dataset$ucgpcode[dataset$ucgpM==0]

            codeset<-sort(codeset)

            num.cases<-length(mdvector)

            random.numbers<-runif(num.cases,0,1)

            iucgp<-dataset$ucgp


            for (i in 1:num.cases)
            {
                  icid<-mdvector[i]
                  impcode<-dataset[icid,164]


                  impcode<-searchcloseby(impcode,codeset)


                        subset<-
dataset$id[is.na(dataset$ucgpcode)==F&dataset$ucgpcode==impcode&dataset
$ucgpM==0]
```

```
                  k<-length(subset)

                  selected<-ceiling(k*random.numbers[i])
                  iucgp[icid]<-dataset[subset[selected],44]

            }

            dataout<-iucgp
            dataout

            }
                  set.seed(100)

      # Occupational class:

      occupimpute<-function(dataset,dataout)
      {
            #dataset<-data

            mdvector<-
dataset$id[dataset$occupM==1&is.na(dataset$occupcode)==F]

            codeset<-dataset$occupcode[dataset$occupM==0]

            codeset<-sort(codeset)

            num.cases<-length(mdvector)

            random.numbers<-runif(num.cases,0,1)

            ioccup<-dataset$occup


            for (i in 1:num.cases)
            {
                  icid<-mdvector[i]
                  impcode<-dataset[icid,165]


                  impcode<-searchcloseby(impcode,codeset)


                        subset<-
dataset$id[is.na(dataset$occupcode)==F&dataset$occupcode==impcode&datas
et$occupM==0]
                  k<-length(subset)

                  selected<-ceiling(k*random.numbers[i])
                  ioccup[icid]<-dataset[subset[selected],31]

            }

            dataout<-ioccup
            dataout

            }
                  set.seed(100)
      # Comorbidity:

      comorbsimpute<-function(dataset,dataout)
      {
            #dataset<-data

            mdvector<-
dataset$id[dataset$comorbsM==1&is.na(dataset$comorbscode)==F]

            codeset<-dataset$comorbscode[dataset$comorbsM==0]

            codeset<-sort(codeset)

            num.cases<-length(mdvector)

            random.numbers<-runif(num.cases,0,1)

            icomorbs<-dataset$COMORBS


            for (i in 1:num.cases)
            {
                  icid<-mdvector[i]
                  impcode<-dataset[icid,166]


                  impcode<-searchcloseby(impcode,codeset)


                        subset<-
dataset$id[is.na(dataset$comorbscode)==F&dataset$comorbscode==impcode&d
ataset$comorbsM==0]
                  k<-length(subset)

                  selected<-ceiling(k*random.numbers[i])
                  icomorbs[icid]<-dataset[subset[selected],14]

            }

            dataout<-icomorbs
            dataout

            }
                  set.seed(100)


      # BMI:

      bmiqimpute<-function(dataset,dataout)
      {
            #dataset<-data

            mdvector<-
dataset$id[dataset$bmiqM==1&is.na(dataset$bmiqcode)==F]
```

```
            codeset<-dataset$bmiqcode[dataset$bmiqM==0]

            codeset<-sort(codeset)

            num.cases<-length(mdvector)

            random.numbers<-runif(num.cases,0,1)

            ibmiq<-dataset$BMIQ


            for (i in 1:num.cases)
            {
                  icid<-mdvector[i]
                  impcode<-dataset[icid,167]


                  impcode<-searchcloseby(impcode,codeset)


                        subset<-
dataset$id[is.na(dataset$bmiqcode)==F&dataset$bmiqcode==impcode&dataset
$bmiqM==0]
                  k<-length(subset)

                  selected<-ceiling(k*random.numbers[i])
                  ibmiq[icid]<-dataset[subset[selected],48]

            }

            dataout<-ibmiq
            dataout

            }
                  set.seed(100)

      # Patient delay:

      ptdlygpimpute<-function(dataset,dataout)
      {
            #dataset<-data

            mdvector<-
dataset$id[dataset$ptdlygpM==1&is.na(dataset$ptdlygpcode)==F]

            codeset<-dataset$ptdlygpcode[dataset$ptdlygpM==0]

            codeset<-sort(codeset)

            num.cases<-length(mdvector)

            random.numbers<-runif(num.cases,0,1)

            iptdlygp<-dataset$ptdlygp


            for (i in 1:num.cases)
            {
                  icid<-mdvector[i]
                  impcode<-dataset[icid,168]


                  impcode<-searchcloseby(impcode,codeset)


                        subset<-
dataset$id[is.na(dataset$ptdlygpcode)==F&dataset$ptdlygpcode==impcode&d
ataset$ptdlygpM==0]
                  k<-length(subset)

                  selected<-ceiling(k*random.numbers[i])
                  iptdlygp[icid]<-dataset[subset[selected],40]

            }

            dataout<-iptdlygp
            dataout

            }
                  set.seed(100)



      #Therapy:

      therapyimpute<-function(dataset,dataout)
      {
            #dataset<-data

            mdvector<-
dataset$id[dataset$therapyM==1&is.na(dataset$therapycode)==F]

            codeset<-dataset$therapycode[dataset$therapyM==0]

            codeset<-sort(codeset)

            num.cases<-length(mdvector)

            random.numbers<-runif(num.cases,0,1)

            itherapy<-dataset$therapy


            for (i in 1:num.cases)
            {
                  icid<-mdvector[i]
                  impcode<-dataset[icid,169]


                  impcode<-searchcloseby(impcode,codeset)
```

```
                        subset<-
dataset$id[is.na(dataset$therapycode)==F&dataset$therapycode==impcode&d
ataset$therapyM==0]
                  k<-length(subset)

                  selected<-ceiling(k*random.numbers[i])
                  itherapy[icid]<-dataset[subset[selected],63]

            }

            dataout<-itherapy
            dataout

            }
                  set.seed(100)


      n.iteration<-50
      nRow <- dim(data)[1]
      cols <- matrix(NA, nRow, 12*n.iteration)

      for (i in 1:n.iteration){

      r<-12*(i-1)+1

      data2<-povgpimpute1(data)
      cols[,r]<-povgpimpute2(data2)
      cols[,r+1]<-totdlygpimpute(data)
      cols[,r+2]<-smkhximpute(data)
      cols[,r+3]<-insgpimpute(data)
      cols[,r+4]<-educimpute(data)
      cols[,r+5]<-ocpimpute(data)
      cols[,r+6]<-ucgpimpute(data)
      cols[,r+7]<-occupimpute(data)
      cols[,r+8]<-comorbsimpute(data)
      cols[,r+9]<-bmiqimpute(data)
      cols[,r+10]<-ptdlygpimpute(data)
      cols[,r+11]<-therapyimpute(data)

            }

      write.csv(cols, file="/Users/Eva/Documents/idata+")
```

APPENDIX D

COMPARISON OF THE DISTRIBUTIONS BETWEEN ORIGINAL UPDATED DATA AND IMPUTED DATA

For total delay:

	Original			Imputed			
	Total (N=320)	White (N= 232)	Black (N=88)		Total (N=490)	White (N=341)	Black (N=149)
Total delay	Frequency (%)	Frequency (%)	Frequency (%)	Total delay	Frequency (%)	Frequency (%)	Frequency (%)
0-<1m	36.99	36.03	39.39	0-<1m	39.39	39.59	38.93
1-<3m	22.25	22.27	22.22	1-<3m	22.04	21.41	23.49
3-<6m	15.03	15.38	14.14	3-<6m	14.29	14.37	14.09
>=6m	25.72	26.32	24.24	>=6m	24.29	24.63	23.49

For smoking history

	Original			Imputed			
	Total (N=320)	White (N= 232)	Black (N=88)		Total (N=490)	White (N=341)	Black (N=149)
Smoking History	Frequency (%)	Frequency (%)	Frequency (%)	Smoking History	Frequency (%)	Frequency (%)	Frequency (%)
never	58.22	54.92	66.36	never	58.57	55.13	66.44
former	30.73	33.71	23.36	former	30.82	34.60	22.15
currrent	11.05	11.36	10.28	currrent	10.61	10.26	11.41

For type of insurance:

	Original			Imputed			
	Total (N=320)	White (N= 232)	Black (N=88)		Total (N=490)	White (N=341)	Black (N=149)
Insurance	Frequency (%)	Frequency (%)	Frequency (%)	Insurance	Frequency (%)	Frequency (%)	Frequency (%)
none	7.28	4.55	14.02	none	9.18	5.87	16.78
public	18.06	9.47	39.25	public	18.57	9.97	38.26
any	74.66	85.98	46.73	any	72.24	84.16	44.97

For education:

	Original			Imputed			
	Total (N=320)	White (N= 232)	Black (N=88)		Total (N=490)	White (N=341)	Black (N=149)
Education	Frequency (%)	Frequency (%)	Frequency (%)	Education	Frequency (%)	Frequency (%)	Frequency (%)
<high school	29.65	19.32	55.14	<high school	31.02	19.65	57.05
high school graduate	32.61	34.85	27.10	high school graduate	34.08	37.83	25.50
>high school	37.74	45.83	17.76	>high school	34.9	42.52	17.45

For oral contraceptive use:

	Original			Imputed			
	Total (N=320)	White (N= 232)	Black (N=88)		Total (N=490)	White (N=341)	Black (N=149)
Oral Contracep-tive use	Frequency (%)	Frequency (%)	Frequency (%)	Oral Contracep tive use	Frequency (%)	Frequency (%)	Frequency (%)
never used	78.44	74.62	87.85	never used	80.82	78.01	87.25
used	21.56	25.38	12.15	used	19.18	21.99	12.75

For usual source of care:

	Original			Imputed			
	Total (N=320)	White (N= 232)	Black (N=88)		Total (N=490)	White (N=341)	Black (N=149)
Usual source of care	Frequency (%)	Frequency (%)	Frequency (%)	Usual source of care	Frequency (%)	Frequency (%)	Frequency (%)
none	11.08	10.27	13.08	none	11.63	10.26	14.77
public	10.54	2.66	29.91	public	11.43	2.05	32.89
private	78.38	87.07	57.01	private	76.94	87.68	52.35

For occupation class:

	Original			Imputed			
	Total (N=320)	White (N= 232)	Black (N=88)		Total (N=490)	White (N=341)	Black (N=149)
Occupation Class	Frequency (%)	Frequency (%)	Frequency (%)	Occupation Class	Frequency (%)	Frequency (%)	Frequency (%)
manager/ profession	23.64	28.63	11.32	manager/ profession	22.24	26.10	13.42
technique/ sales	29.62	37.02	11.32	technique/ sales	30.41	39.00	10.74
skilled worker	20.38	14.50	34.91	skilled worker	21.02	14.66	35.57
unskilled worker	12.5	6.11	28.30	unskilled worker	12.04	5.28	27.52
homemaker	13.86	13.74	14.15	homemaker	14.29	14.96	12.75

For comorbidities:

	Original			Imputed			
	Total (N=320)	White (N= 232)	Black (N=88)		Total (N=490)	White (N=341)	Black (N=149)
Co-morbidity	Frequency (%)	Frequency (%)	Frequency (%)	Co-morbidity	Frequency (%)	Frequency (%)	Frequency (%)
no	28.31	34.54	14.89	no	29.59	35.78	15.44
yes	71.69	65.46	85.11	yes	70.41	64.22	84.56

For BMI:

	Original			Imputed			
	Total (N=320)	White (N= 232)	Black (N=88)		Total (N=490)	White (N=341)	Black (N=149)
BMI	Frequency (%)	Frequency (%)	Frequency (%)	BMI	Frequency (%)	Frequency (%)	Frequency (%)
low normal	25.6	33.33	5.98	low normal	25.31	34.02	5.37
high normal	19.32	23.23	9.40	high normal	18.78	22.87	9.40
over-weight	23.19	22.22	25.64	over-weight	23.88	21.99	28.19
very over-weight	31.88	21.21	58.97	very over-weight	32.04	21.11	57.05

For patient delay:

	Original			Imputed			
	Total (N=320)	White (N= 232)	Black (N=88)		Total (N=490)	White (N=341)	Black (N=149)
Patient Delay	Frequency (%)	Frequency (%)	Frequency (%)	Patient Delay	Frequency (%)	Frequency (%)	Frequency (%)
no	6.98	8.63	2.91	no	5.92	7.62	2.01
0-<1m	59.78	60.78	57.28	0-<1m	61.02	61.58	59.73
1-<3m	13.41	12.94	14.56	1-<3m	12.86	11.73	15.44
3-<6m	8.1	7.45	9.71	3-<6m	7.76	7.33	8.72
>=6m	11.73	10.20	15.53	>=6m	12.45	11.73	14.09

For treatment:

	Original			Imputed			
	Total (N=320)	White (N= 232)	Black (N=88)		Total (N=490)	White (N=341)	Black (N=149)
Treatment	Frequency (%)	Frequency (%)	Frequency (%)	Treatment	Frequency (%)	Frequency (%)	Frequency (%)
no surgery	12.53	5.35	27.86	no surgery	12.04	5.57	26.85
only surgery	43.28	49.50	30.00	only surgery	44.49	49.85	32.21
surgery+	44.19	45.15	42.14	surgery+	43.47	44.57	40.94

BIBLIOGRAPHY

American Cancer Society. *Cancer Facts & Figures: 1983*. New York: American Cancer Society Inc., 1984.

Barrett II RJ, Harlan LC, Wesley MN, Hill HA, Chen VW, Clayton LA, Kotz HL, Eley JW, Robboy SJ, and Edwards BK. *Endometrial Cancer: Stage at Diagnosis and Associated Factors in Black and White Patients.* American Journal of Obstetrics and Gynecology 173(2): 414-423, 1995.

Coates RJ, Click LA, Harlan LC, Robboy S, Barrett II RJ, Eley JW, Reynolds R, Chen VW, Darity WA, Blacklow RS, and Edwards BK. Differences between Black and White Patients with Cancer of The Uterine Corpus in Interval from Symptom Recognition to Initial Medical Consultation (United States). Cancer Causes & Control 7: 328-336, 1996.

Cox DR. *Regression models and life-tables*. Journal of the Royal Statistical Society. 34(2): 187-220, 1972.

Cohen J. *A coefficient of agreement for nominal scales*. Educational and Psychological Measurement 20(1): 37-46, 1960.

Dolinsky C. *Endometrial Cancer: The Basics.* http://www.oncolink.org/types/article.cfm?c=6&s=18&ss=137&id=8227. 2008.

Dutton DB. Hematocrit levels and race: An argument against the adoption of separate standards in screening fro anemia. J Natl Med Assoc 71: 945-954, 1979.

Edwards BK. *SEER Cancer Statistics Review, 1975-2005, National Cancer Institute.* Bethesda, Maryland, http://seer.cancer.gov/csr/1975_2005/, based on November 2007 SEER data submission, posted to the SEER web site, 2008.

FIGO: Annual report of the results of treatment in gynecological cancer. Int J Gynecol Obstet 20:75-7, 1988.

Hill HA, Eley JW, Harlan LC, Greenberg RS, Barrett II RJ, Chen V. *Racial differences in endometrial cancer survival: the black/white cancer survival study.* Obstetrics & Gynecology 88: 919-926, 1996.

Howard J, Hankey BF, Greenberg RS, Austin DF, Correa P, Chen VW, Durako S. A collaborative study of differences in the survival rates of black patients and white patients with cancer. Cancer 69: 2349-60, 1992.

Little RJA, Rubin DB. *Statistical analysis with missing data.* New York: John Wiley & Sons, 1986.

Miller BA, Ries LAG, Hankey BF. *SEER cancer statistics review: 1973-1990.* Bethesda, Maryland: National Cancer Institute: National Institutes of Health. Report No: NIH-NCI-93-2789, 1993.

Myers MH, Hankey BF. *Cancer Patient Survival Experience.* National Institutes of Health publication No. 80-2148. 1980.

Myers MH, Hankey BF. *Cancer sites for inclusion in the study of black/white survival differences.* Unpublished memorandum, National Cancer Institute, 1984.

Ries LAG, Melbert D, Krapcho M, Stinchcomb DG, Howlader N, Horner MJ, Mariotto A, Miller BA, Feuer EJ, Altekruse SF, Lewis DR, Clegg L, Eisner MP, Reichman M, Ries LG, Pollack ES, Young JL Jr. *Cancer patient survival: Surveillance, Epidemiology, and End Results Program, 1973-79.* J Natl Cancer Inst 70: 693-707, 1983.

Hill HA, Coates RJ, Austin H, Correa P, Robboy SJ, Chen V, Click LA, Barrett II RJ, Boyce JG, Kotz HL, and Harlan LC. *Racial Differences in Tumor Grade Among Women with Endometrial Cancer.* Gynecologic Oncology 56: 154-163, 1995