APPLICATION OF MULTIPLE IMPUTATION IN ANALYSIS OF MISSING DATA IN A STUDY OF HEALTH-RELATED QUALITY OF LIFE

by

Chunming Zhu

B.S. of Biological Science & Biotechnology, Tsinghua University, China, 1994

Ph.D of Biophysics, Tsinghua University, China, 1999

Submitted to the Graduate Faculty of

Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Master of Science

University of Pittsburgh

2011

UNIVERSITY OF PITTSBURGH

Graduate School of Public Health

This thesis was presented

by

Chunming Zhu

It was defended on

March 30, 2011

and approved by

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

Committee Member:
Lan Kong, PhD
Assistant Professor
Department of Biostatistics
Graduate School of Public Health
University of Pittsburgh

Committee Member:
Greg Yothers, PhD
Research Assistant Professor
Department of Biostatistics
Graduate School of Public Health
University of Pittsburgh

Committee Member:
Tianjiao Chu, PhD
Assistant Professor,
Department of Obstetrics, Gynecology & Reproductive Sciences
School of Medicine
University of Pittsburgh

Gong Tang, Ph.D

APPLICATION OF MULTIPLE IMPUTATION IN ANALYSIS OF MISSING DATA IN A STUDY OF HEALTH-RELATED QUALITY OF LIFE

Chunming Zhu, MS

University of Pittsburgh, 2011

When a new treatment has similar efficacy compared to standard therapy in medical or social studies, the health-related quality of life (HRQL) becomes the main concern of health care professionals and can be the basis for making a decision in patient management. National Surgical Adjuvant Breast and Bowel Protocol (NSABP) C-06 clinical trial compared two therapies: intravenous (IV) fluorouracil (FU) plus Leucovorin (LV) and oral uracil/ftorafur (UFT) plus LV, in treatment of colon cancer. However, there was a high proportion of missing values among the HRQL measurements that only 481 (59.8%) UFT patients and 421 (52.4%) FU patients submitted the forms at all time points. Ignoring the missing data issue often leads to inefficient and sometime biased estimates.

The primary objective of this thesis is to evaluate the impact of missing data on the estimated the treatment effect. In this thesis, we analyzed the HRQL data with missing values by multiple imputation. Both model-based and nearest neighborhood hot-deck imputation methods were applied. Confidence intervals for the estimated treatment effect were generated based on the pooled imputation analysis.

The results based on multiple imputation indicated that missing data did not introduce major bias in the earlier analyses. However, multiple imputation was worthwhile since the most estimation from the imputation datasets are more efficient than that from incomplete data.

These findings have public health importance: they have implications for development of health policies and planning interventions to improve the health related quality of life for those patients with colon cancer.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

It is never enough to state my sincere gratitude to my thesis advisor, Dr. Gong Tang. I must say that missing data analysis is one of the most difficult topics I have learned in statistics. But he helped me go through all the difficult times and give me an invaluable learning and research experience. Many thanks to Dr. Lan Kong, who is my academic advisor since I entered the Department of Biostatistics. She gave me a lot of help and wise advice in last two years. I also would like to thank Dr. Greg Yothers, who is a member of the NSABP Biostatistical Center and kindly shared these important data with me for this thesis, carefully read my thesis and made several corrections. Many thanks to Dr. Tianjiao Chu, who gave me opportunity to work in his lab as statistics staff where I gained hands-on experience in data management, statistical analysis and programming techniques.

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. I also like to give thanks to my classmates and friends for their help and support.

Much gratitude is owed to my family: my wife and two lovely children, their love and encouragement is one of the most important motivation and strength for me to study and work hard.

Finally, I would like to say: Hallelujah!

1.0 INTRODUCTION

In clinical studies that compare one or more new treatments with a standard treatment, the efficacy in clinical outcomes such as patient survival and clinical response is usually the basis for treatment selection decision. However, in many circumstances, a new treatment may not demonstrate superiority in efficacy over the standard treatment. When efficacy is similar, other factors have to be incorporated in the decision making. Health-related quality of life (HRQL) outcomes often become an important indicator in treatment benefit in those scenarios and healthcare professionals may rely on them to select appropriate treatments for their patients. With advances in medical science and technology, there are an increasing number of people living with chronic diseases and disabilities. Long-term HRQL is useful for evaluating the benefit of a treatment beyond its efficacy in clinical outcomes. The change in our population's morbidity pattern has called for a paradigm shift in how we should evaluate outcomes of illness and care.

Colon cancer is the fourth most common form of cancer in the United States and the third leading cause of cancer-related death in the Western world. Invasive cancers that are confined within the wall of the colon (stages I and II) are often curable with surgery alone. If untreated, they may spread to regional lymph nodes (stage III), where up to 73% are curable by surgery and chemotherapy. Cancer that metastasizes to distant sites (stage IV) is usually not curable, although chemotherapy can extend survival. In rare cases, surgery and chemotherapy together have seen patients through to a cure. In patients in stage II and stage III colon cancer, adjuvant therapy of surgery and chemotherapy is the most common treatment. Chemotherapy with intravenous fluorouracil (FU) and leucovorin (LV) has been demonstrated to prolong disease-free survival (Wolmark, 1993). Another approach of chemotherapy is oral administration of uracil/ftorafur (UFT) plus leucovorin (LV). A small phase II study in Taiwan indicated that this new approach could be comparable to intravenous (IV) fluorouracil (FU) and leucovorin (LV) in stage IV colorectal cancer (Yang, 2002).

Adverse events associated both of these two treatments were reported. These two regimens were associated with widely different acute and late effects, which could be both physical and psychological in nature. The most common adverse reactions were GI toxicity (diarrhea, nausea, and stomatitis) and granulocytopenia (Wolmark, 1998). HRQL concerns are therefore important for these patients for selecting between the different treatment options. In the NSABP C-06 trial, not only the disease-free survival and overall survival were compared between these two regimens, but also the health-related quality of life outcomes from patients were studied.

1.1 THE QUALITY OF LIFE STUDY IN NSABP C-06 TRAIL

National Surgical Adjuvant Breast and Bowel Protocol (NSABP) C-06 trail was a randomized equivalence trial to compare the intravenous FU plus LV with the oral UFT and LV. In this trial, patients were randomly assigned to either FU arm or UFT arm. Those assigned to FU+LV received LV 500mg/m^2 by IV infusion over 2 hours and FU 500mg/m^2 by IV bolus

1hour after LV infusion weekly for 6 weeks, followed by a rest period. Treatment was restarted 21 days after the date of administration of the sixth dose of the previous cycle (1 cycle = 8 weeks). A total of three cycles were administrated. Patients assigned to UFT+LV received UFT 300mg/m^2/day plus LV 90mg/day for 28 days followed by a 7-day rest period. Patients in this group took both drugs orally, the total daily dose divided into three doses to be taken 8 hours apart. A total of five cycles were administrated. Chemotherapy in both arms began within 1 week from randomization and 7 weeks from surgery. Disease free survival and overall survival data were recorded. During chemotherapy and after 1 year follow up, measurement of Health-related quality of life (HRQL) was carried out. The study has demonstrated that two regimens are equivalent in terms of disease-free survival and overall survival (Lembersky 2006).

Health-related quality of life was measured with Functional Assessment of Cancer Therapy-Colorectal (FACT-C) questionnaire, Short form-36 Vitality Scale (SF-36), and Quality of life Rating Scale (QLRS) at baseline, once during chemotherapy (16weeks in FU arm and 15 weeks in UFT arm), and at 1 year.

Functional Assessment of Cancer Therapy-Colorectal (FACT-C) questionnaire is one of the two quality of life (QOL) assessment tools available for colorectal cancer patients. The FACT-C combines specific concerns related to colorectal cancer with concerns that are common to all cancer patients. It is a multi-dimensional, 44-item, cancer-oriented measure. Six subscales provided scores for physical well-being (pwb), social/family well-being (swb), relationship with physician (rwd), emotional well-being (ewb), functional well-being (fwb), and problems commonly experienced by patients (fc) with colorectal cancer.

The SF-36 was developed from work done by the RAND Corporation and the Medical Outcomes Study (MOS), based on the measurement strategy of the RAND Health Insurance

Study in the 1980s. SF-36 is a multi-purpose, short-form health survey with only 36 questions. It yields an 8-scale profile of functional health and well-being scores as well as psychometrically-based physical and mental health summary measures and a preference-based health utility index. The SF-36 can be either self-administered or administered by a trained interviewer, either in person or by telephone. Over the years, the SF-36 has been used in surveys of general and specific populations, for comparing the relative burden of diseases across different sub-groups and in differentiating the health benefits produced by health care treatments.

The Quality of Life Rating Scaling (QLRS) evaluates the patients overall perception of quality of life on a 0 to 10 scale, where 0 indicates the lowest and 10 indicates the highest possible quality of life. Higher scores on all of three measures indicate better health-related quality of life.

1.2 THE ISSUE OF MISSING DATA

The NSABP C-06 study randomized 1608 patients: 803 to the FU arm and 805 to the UFT arm. The patients who contributed to the HRQL analysis were similar to the full study population (see Section **3.1**). Almost 60% were 60 years old or older, slightly over 50% were male, and 78% were white. The distribution of patient and tumor characters was similar in both arms of the trial.

The major reason for incomplete records in this study is loss to follow up. Among the patients participating in the HRQL study, 79% of the patients in UFT arm and 73% in the FU arm completed the FACT-C at week 15/16 (during chemotherapy), while 67% and 62%, completed it at 1 year, respectively. A total 481 (59.8%) UFT patients versus 421(52.4%) FU patients submitted the forms containing QLRS and SF36 vitality scales at all time points. Since there is a difference in the response rates between two treatment arms (Kopec, 2007), data might not be missing completely at random. Analyses that are solely based on complete cases, which had complete records in the HRQL outcomes often lead to inefficient and sometimes biased estimates.

1.3 STUDY OBJECTIVES

The main objective of this thesis is to explore the impact of missing-data. The aim of imputing missing-data is 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 the difference (or similarity) of HRQL for the two treatment arms. A second objective is to estimate the variance of parameter estimates derived from imputed datasets.

In this thesis, multiple imputation approaches were applied to handle the missing data. Through multiple imputation, the missing values were filled in according to an appropriate algorithm. Subsequently standard analysis techniques were performed on the resulted complete datasets. To account for the variation of imputed values, more than one value (for example, 20) were filled in for each missing value so that multiple datasets were generated for complete-data analysis. Then the estimates were compared to that from the original incomplete dataset. Variability of parameter estimates within each imputed dataset was also estimated.

2.0 REVIEW OF METHODS FOR ANALYSIS OF MISSING DATA

Missing data are prevalent in data collected throughout various scientific fields. Missing data may occur in two different formats: no information is provided on one or more items or no information is available from a whole unit. The former case is called item nonresponse and the latter one is called unit nonresponse. Some items may tend to have more missing values than others. For example, income is often an important variable in survey studies but many surveyed subjects tend not to report their actual income or income level because of privacy concern. Dropout is a typical missingness that occurs in studying the development of a process over time. In such studies one or more outcome variables are repeatedly measured over a certain period of time. Missing values appear when a participant drops out prematurely before the end of the study and all measurements of the outcome variable(s) are missing after that occasion. Missing values occur for a variety of reasons; such as that collection on a portion of subjects is discontinued after certain period by design, some subjects become non-compliant, or data collection is not done properly.

Rubin (1976) introduced a classification system on missing-data mechanisms and it is widely used in the literature. Depending on how the missing data process is related to the underlying hypothetical complete data, three missing-data mechanisms were defined by Rubin (1976). The data are called missing completely at random (MCAR) if the missingness of a variable Y is unrelated to either the value of Y or that of other measured variables. In other words, the observed data points are a simple random sample of data had the data been complete. Data are called missing at random (MAR) when the missingness of a variable Y is unrelated to the value of Y itself after conditioning on other observed values. Finally, data are called missing not at random (MNAR) when the missingness of a variable Y still depends on the value of Y even given the observed values.

In principle, it is possible to verify whether the data are MCAR or not. It is impossible to test the MAR mechanism except under certain parametric framework when the missing-data mechanism is modeled by a parametric model. This is an important practical problem for missing data analysis because both of the two popular techniques for analysis of missing data, the ignorable maximum likelihood method and multiple imputation, assume an MAR mechanism.

2.1 COMPLETE CASE ANALYSIS

List-wise and pair-wise deletion methods are by far the most prevalent approaches in analysis of missing data in practice. The list-wise deletion approach removes the variables with missing data from the inferential procedure and the pair-wise deletion approach removes subjects with missing values (Enders, 2010). The advantage of these methods is that they are convenient and are standard options in statistical software packages. However, list-wise deletion may remove variables of interest and the pair-wise deletion assumes MCAR data and can produce distorted parameter estimates when this assumption does not hold. Even if the MCAR assumption is plausible, eliminating data leads to inefficient estimates. Consequently, these are not recommended unless the portion of missing data is very small.

2.2 LIKELIHOOD-BASED METHODS

The maximum likelihood method (ML) (Little and Rubin, 2002) maximizes the likelihood based on ($Y_{i,obs,}$ R_i). Denote X={x_i}$_{i=1,2,\ldots n}$, the covariates, Y={y_i}$_{i=1,2,\ldots n}$, the outcome, R={R_i}$_{i=1,2,\ldots n,}$ the missing-data indicator, θ is parameter of interest, and φ is unknown parameter that related to the missingness of data. The likelihood function:

$$\mathrm{L}(\theta,\varphi\ |\mathrm{X};\mathrm{Y_{obs}};\mathrm{R})\ \propto \prod_{i=1}^{n} p\big(y_{i,obs,}R_i\big|\ \mathrm{x_{i;}}\ \theta,\varphi\big)$$

$$= \prod_{i=1}^{n} \int p\big(y_{i,obs,}y_{i,mis,}R_i\big|\ \mathrm{x_{i;}}\ \theta,\varphi\big)\ dy_{i,mis}$$

$$= \prod_{i=1}^{n} \int p\big(y_{i,obs,}y_{i,mis}\big|\ \mathrm{x_{i;}}\ \theta\big)\ p(R_i|\ \mathrm{x_{i;}}\ y_{i,obs,}\ y_{i,mis};\varphi)dy_{i,mis}$$

When the data are MAR,

$$p(R_i|\ \mathrm{x_{i;}}\ y_{i,obs,}\ y_{i,mis};\varphi) = p(R_i|\ \mathrm{x_{i;}}\ y_{i,obs}\ ;\varphi) \text{ and}$$

$$\mathrm{L}(\theta,\varphi\ |\mathrm{X};\mathrm{Y_{obs}};\mathrm{R})\ \ \propto\ \ \mathrm{L}(\varphi|R_i,\mathrm{x_{i,}}y_{i,obs})\ \mathrm{L}(\theta|\mathrm{X};\mathrm{Y_{obs}};\mathrm{R})$$

Where,

$$\mathrm{L}(\theta|\mathrm{X};\mathrm{Y_{obs}};\mathrm{R}) = \prod_{i=1}^{n} p\big(y_{i,obs}\big|\ \mathrm{x_{i;}}\ \theta\big) \text{ is the ignorable likelihood and}$$

$$\mathrm{L}(\varphi|R_i,\mathrm{x_{i,}}y_{i,obs}) = \prod_{i=1}^{n} p(R_i|\ \mathrm{x_{i;}}\ y_{i,obs};\varphi)$$

is only related to missing-data mechanism. If θ and φ are also distinct, the inference on θ does not depend on the missing-data mechanism. Therefore when data are MAR, and θ and φ are distinct, the missing-data mechanism is ignorable. When data are MNAR, ignoring missing-data mechanisms could lead to biased estimates of θ. In such circumstances, a parametric form has to be assumed for the missing-data mechanism in the ML method. And the inference can be highly sensitive to such assumptions.

2.3 ESTIMATION EQUATION-BASED METHODS

Inverse probability weighted estimating equations (IPWEE) is an estimating equation based method (Robins, Rotnitzky and Zhao, 1995) to make inference on selection models. A simple version of this method is to weigh each complete case by the inverse probability of being observed while constructing the estimating equation. The motivation is that each complete case not only represents itself but also other incomplete cases with similar characteristics. It still requires specifying a model for the missing-data mechanism. Mis-specification often leads to biased estimates for the model parameters.

2.4 MULTIPLE IMPUTATION

Imputation is a general and flexible method for analysis of missing data . Missing values are imputed by draws from a predictive distribution of the missing values based on observed data. Such a predictive distribution could be an explicit model, such as unconditional mean, conditional mean or stochastic imputation or an implicit model such as hot deck, substitution, or cold deck imputation (Little and Rubin, 2002).

Imputations should generally be:

a) Conditional on observed variables, to reduce bias due to no response, improve precision, and preserve association between missing and observed variables;

b) Multivariate, to preserve association between missing variables;

c) Draws from a predictive distribution rather than means, to provide valid estimates of a wide range of model parameters.

The limitation of single imputation is when a value filled to the missing value, we assume that the filled-in value is a true observed value and the data are complete. But the fact is that we do not know the true value of the missing value. In other words, the filled-in value itself has some uncertainty and it is not appropriate to assume it fixed. To account for this variability across imputations, we can create multiply imputed data sets that allow the additional uncertainty from imputations to be assessed. In other word, multiple imputations overcome the important limitation of single imputation, where standard variance formulas applied to a single imputed dataset systematically underestimate the variance of estimates, even when the model used to generate the imputations is correct (Little and Rubin, 2002). In this thesis, two multiple imputation methods, explicit modeling based and hot deck methods, are used to generate imputed data sets from the original HRQL data. Details on these two methods are introduced in Chapter 3.

3.0 METHODS

3.1 IDENTIFY MISSING DATA AND PREPARE THE HRQL DATASET

3.1.1 Identify missing-data patterns

In the NSABP C-06 trial, 780 out of 803 patients in the FU arm and 784 of 805 in the UFT arm participated in the HRQL study. Among the patients in FU arm, 413 patients completed all information of the FACT-C, QLRS and SF36 vitality scales forms. In UFT arm, 468 had complete information. Using R package *mice* (multivariate imputation by chained equations), missingness indicator matrices were generated for both FU and UFT data sets. According to the indicator matrices, there are 19 missing patterns in FU data set and 25 patterns in UFT data set. To simplify the imputation, we deleted those missing-data patterns that had only 1 or 2 subjects. After all we had 766 patients with seven missing-data patterns in the FU arm and 761 patients with six missing-data patterns in the UFT arm (**Figure 1**). The datasets from two arms were kept separated for imputation approaches. Meanwhile, a data set with missing values was generated by combining the data from two arms.

3.1.2 Check difference between complete and incomplete observations

Most of the missingness is from dropout. The dropouts include 268 (139+129) out of 766 patients (89%) in the FU arm and 225 (131+94) out of 761 patients in the UFT arm. Combining data from the dropouts with those from the complete cases forms a monotone missing-data pattern. To check if the missingness is related to the patients score or/and the treatment, we draw trend plots of mean values of each items in the FACT-C, QLRS and sf-36 forms versus the time points (**Figure 2**).

3.2 IMPUTATION APPROACHES

It was suspected that the missingness was mostly associated with the treatment and the data were missing at random, so we can use multiple imputation approaches. The multiple imputation method basically can be carried out by three steps: imputation step, analysis step and summary steps. In the imputation step, each missing value was filled with several values (in this thesis, 20) based on a predictive distribution based on the observed data and 20 complete data sets are generated. We used two approaches in the imputation step, an explicit model-based imputation method and a nearest-neighborhood hot-deck imputation.

In analysis step, each imputed data set was analyzed separately, including:

a) Check the descriptive statistics for each variable in the 20 imputed datasets, such as mean, standard errors;
b) Build the models with selected covariates in each of the 20 imputed datasets.
c) Calculate the estimated coefficients (θd), d=1, …, 20 and standard errors ($\widehat{SE}_d$), d=1,…, 20 for treatment in the models, and within imputation variance is $\widehat{W}_d$= $(\widehat{SE}_d)^2$, d=1,…, 20.

In the summary step, the analysis results from each imputed dataset were combined to obtain the final statistical inference. The final parameter estimates were the average of corresponding parameter estimates that were resulted from 20 imputed datasets. The variation of the final parameter estimates were estimated by the sum of between-imputation variation and within-imputation variation. The between-imputation variance can be estimated by:

$$B_{20} = \frac{1}{20-1}\sum_{d=1}^{20}(\hat{\theta}_d - \bar{\theta}_{20})^2,$$

where $\bar{\theta}_{20} = \frac{1}{20}\sum_{d=1}^{20}\hat{\theta}_d.$

And the total variability associated with the imputation of the missing values can be estimated by:

$$T_{20} = \bar{W}_{20} + \frac{20+1}{20} B_{20}$$

Where $\bar{W}_{20} = \frac{1}{20}\sum_{d=1}^{20}\widehat{W}_d = \frac{1}{20}\sum_{d=1}^{20}(\widehat{SE}_d)^2.$

At the final step estimate the fraction of information about treatment effect missing due to incomplete data by:

$$\hat{\gamma}_{20} = (1+\frac{1}{20})\, B_{20}/T_{20}$$

3.2.1 Explicit model-based Imputation

Consider a dataset where $Y_1,\ldots.,Y_{k-1}$ is observed and Y_k has missing values. Then we can impute a conditional draw based on a regression model:

$$\hat{y}_{ik} = \tilde{\beta}_{K0.1.2\ldots K-1} + \sum_{j=1}^{K-1}\tilde{\beta}_{Kj.1.2\ldots K-1}\, y_{ij} + z_{iK}$$

Where $\mathbf{z_{ik}}$ is a random normal deviate with mean 0 and residual variance in the regression of Y_k on $Y_1,\ldots.Y_{k-1}$ based on the complete cases. The addition of the random normal deviate makes the imputation a draw from the predictive distribution of the missing values. In above linear regression model, we assume the predictive distribution of missing values is normal distributed. When this normality assumption is not appropriate, we may use other models to approximate.

For example, if Y_k is binary, we can use logistic model to approximate. And the random draws are from the complete cases with similar propensity of missingness.

In our HRQL data, the Sf-36 score has a scale from 0 to 100. The histogram (not shown) of the scores indicates that a normal distribution is appropriate. The relationship with physician (rwd) score in FACT-C, NRMACT (return to normal action) score and QLRS score only have fewer 10 or less levels and are highly skewed. It would be inappropriate to assume that they are normally distributed. We can keep these variables as categorical, and use a *linear discriminant analysis* (lda) method to impute the missing values. While the physical well-being (pwb), social/family well-being (swb), emotional well-being (ewb), functional well-being (fwb), and problems commonly experienced (fc) score has a scale from 0 to 28. If we consider them as categorical variables, there are too many levels which cause severe collinearity when we try to use the *mice* package for imputation. So we assume that they are continuous and use Bayesian linear regression to impute the missing values in them as in the Sf36 scores. This imputation approach can be done in the R package *mice,* by setting imputation method to be “lda” for rwd, NRMACT and QLRS, and “norm”, which means Bayesian linear regression, for Sf-36 scores and other variables.

3.2.2 Hot deck Imputation

As described before, hot deck imputation is a method based on an implicit model. With most hot deck procedures, missing values are replaced by values from similar responding units in the sample. It could be simply random sampling from the observed values with replacement, or random sampling within adjustment cells of observed values.

A more general 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 $x_{i1}, \ldots, x_{ij}$ 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

$$d\,(i,\ i') = \sum_{j=1}^{J} |\ x_{ij} - x_{i'j}|$$

We might choose an imputed value for y_i from those unit i' that are such that

(1) $y_{i'}, x_{i'1}, \ldots, x_{i'j}$ 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, 2002)

In our HRQL dataset the unit is an individual case with its corresponding scores as the covariates. We consider those units within the same missing-data pattern as a pool, so the programming can be more proficient. The procedure consists of two stages. First, measure the distances between a unit with missing values and units from complete cases and choose complete units which have the closest distances to the unit as its nearest neighborhood. Then a value is randomly drawn from the corresponding variable in its neighborhood to insert in place of the missing value. If the unit have more than one missing values, repeat the random draw, until all the missing values are filled in this unit. Then the imputation process described above is repeated *n* times to create *n* complete data sets. These *n* datasets are analyzed separately and the results are combined to form one overall inference.

It is very important to define an appropriate metric to measure distance between units. In our dataset, NRMACT, QLRS have levels 0-10, *rwd* has levels 0-8, sf36 score has levels 0-100, and other scores have 28 levels (0-28). So one approach is to define a metric with normalizing all variables to have same number of levels (0-10), which can be written as:

$$d(i, i') =[\sum_{j=1}^{J} | x_{ij} - x_{i'j}| + \frac{10}{8} | x_{ik} - x_{i'k}| + \frac{10}{28} \sum_{l=1}^{L} | x_{il} - x_{i'l}| + \frac{10}{100} | x_{im} - x_{i'm}|] / N_{obs}$$

Where j indicate the variable NRMACT and QLRS; k, rwd; m, 36sf; l for other variables; and N_{obs} is the number of observed values in that subject. There are 9 variables in the data, and each variable has three time points. If a subject didn't fill out the form at last time point (1 year) and complete the other two time points then the N_{obs} for this subject is 9*2=18. This adjustment makes the distances from all patterns comparable.

3.3 ANALYSIS HRQL WITH IMCOMPLETE AND IMPUTED DATA

An important end point for the HRQL study was the FACT-C total score. It was the normalized sum of scores from the six subscale scores, physical well-being, social/family well-being, relationship with physician, emotional well-being, functional well-being, and problems commonly experienced. The normalized score is in scale 0 to 100, with higher score indicates better quality of life. The QLRS and sf-36 scores can be used for analysis directly.

One approach is to compare the treatment effect on HRQL by comparing the last time point scores (at 1 year), including FACT-C total scores, QLRS and SF 36 scores, using linear regression models with treatment and the baseline scores as predictors. This approach is intuitive and easy to use and is performed in STATA. The linear regression models can be expressed as:

$$Y_1 = \beta_0 + \beta_1 * trt + \beta_2 * Y_0,$$

where Y_0 and Y_1 are scores at the baseline and 1 year, respectively.

Another Statistical comparison is using GEE (generalized estimation equations) models with time points as the repeated variable, an unstructured covariance pattern, robust variance estimator were used in this model. This method utilizes all available values even there is missing value in some observations. This approach is also carried out in STATA, using *xtgee* procedure. All of above approaches were used for all the imputed data sets and the original incomplete data set.

4.0 RESULTS

4.1 MISSINGNESS OF THE DATA

4.1.1 Missing patterns

In the updated dataset, we have 766 patients with 7 missing patterns in 5FU arm and 761 patients with 7 patterns in UFT arm (**Figure 1**).

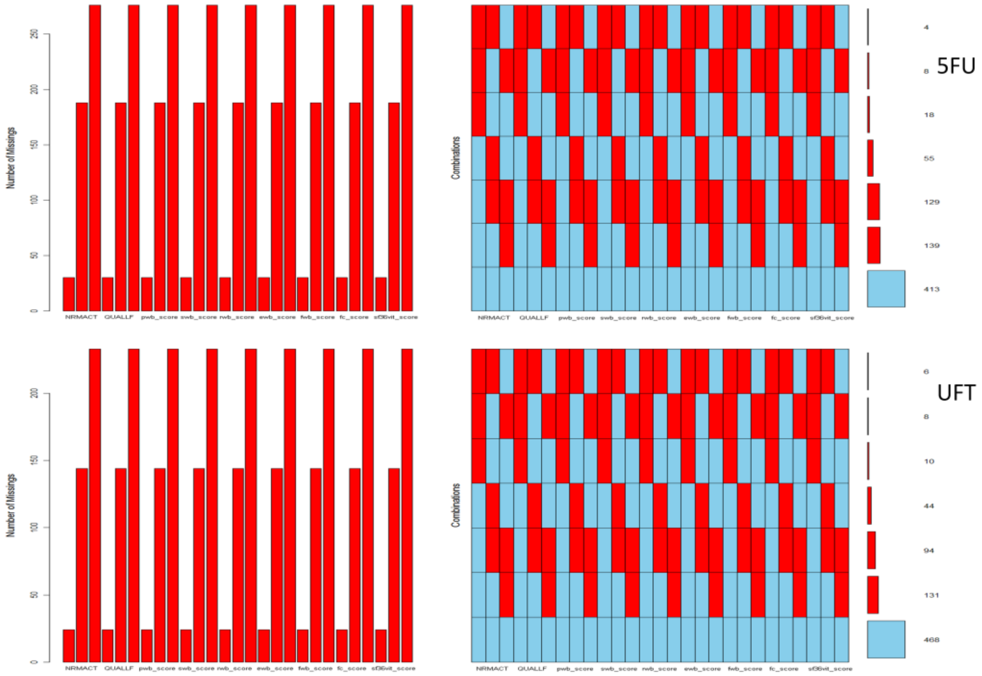

Figure 1. missing pattern of the data.

In **Figure 1**, the top panel is 5FU arm, and lower panel is UFT arm. The left shows the number of missing values for each variable at different time points. The higher of the red bar means more missing values. In the right, missing patterns are showed as two color blocks, in which blue block means the values in the block were observed and red means the values were missing. We noticed that when a patient did not have a response in one score at one time point, there is no record of all variables at that time point. Which means the patient did not turn in the questionnaires form or fill the form at that time point. When a patient filled the form at a time point, he or she answered all the questions. About 35.0% in 5FU arm and 29.6% in UFT arm were drop-out, either during chemotherapy or at 1-year. Together with the complete cases, they form a monotonic missing pattern. But there are about 10 percent of incompleteness does not belong to monotonic missing pattern, in which 85 out of 766 in FU arm (11%) and 68 out of 761 in UFT arm (8.9%).

4.1.2 Difference in drop-out between treatments

To check if the drop-out is related to the score of those patients, we draw trends of score for complete cases and those with dropout (**Figure 2**). In **Figure 2,** the y axis the scores of each covariate, the x axis is the time point. The solid lines are for 5FU arm and dash lines for UFT arm. Red color is for drop out observations. For all the complete cases, there is no significant difference between two treatment arms. For drop out cases, the mean observed values are lower than those complete cases for variable NRMACT, QLRS, 36sf and most of FACT-C scores. The mean score difference between complete case and dropout case are not statistically significant. Therefore there is no obvious evidence against an MCAR mechanism.

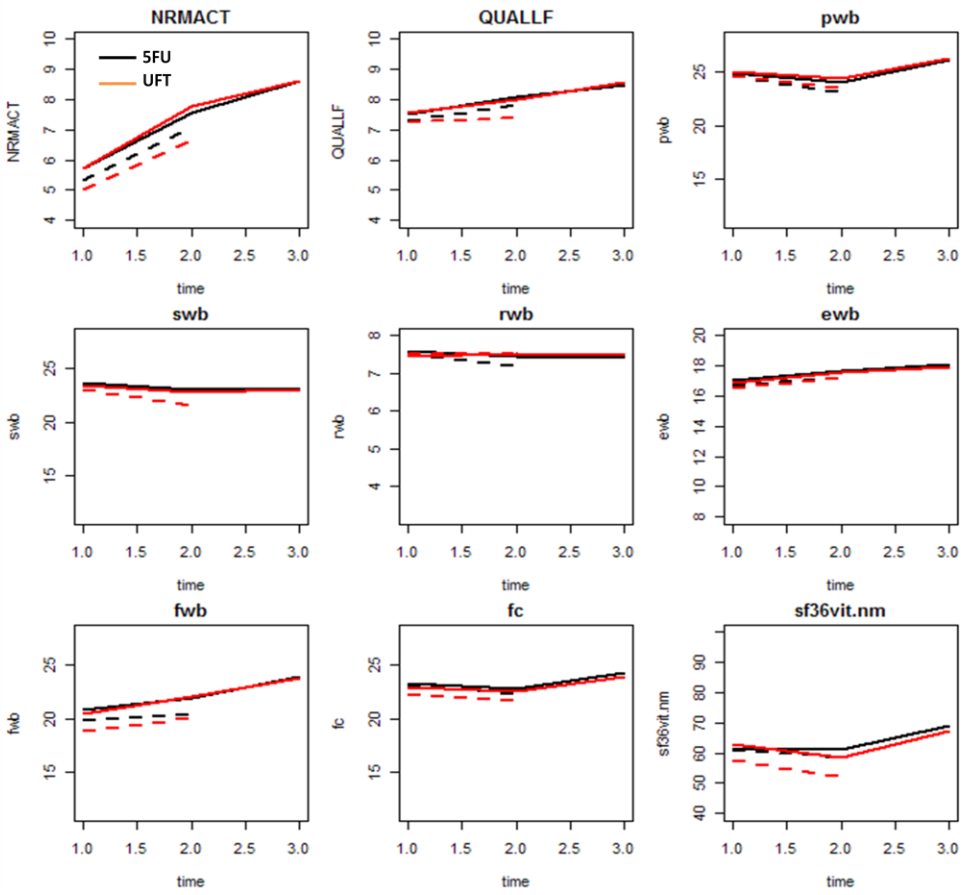

Figure 2. Comparison of the complete cases and the dropout cases in FU and UFT arms.

4.2 COMPARE THE INCOMPLETE AND IMPUTED DATASETS

To compare the imputed datasets to the original incomplete dataset, the mean values of the original incomplete dataset are summarized in **Figure 3** and **Table 1.**

Table 1. Summary of the dataset before imputation

	FACTC_n_score			QLRS			Sf36 Vitality		
	baseline	chemo	1-year	baseline	chemo	1-year	baseline	chemo	1-year
FU.mean	82.99	82.93	87.16	7.45	8.04	8.45	61.03	60.41	68.02
FU.sd	11.23	11.66	11.37	1.96	1.83	1.87	22.37	22.00	22.22
UFT.mean	82.10	82.56	86.78	7.49	7.88	8.53	61.25	57.15	66.86
UFT.sd	11.26	12.13	10.87	1.80	1.87	1.70	21.88	23.02	21.00

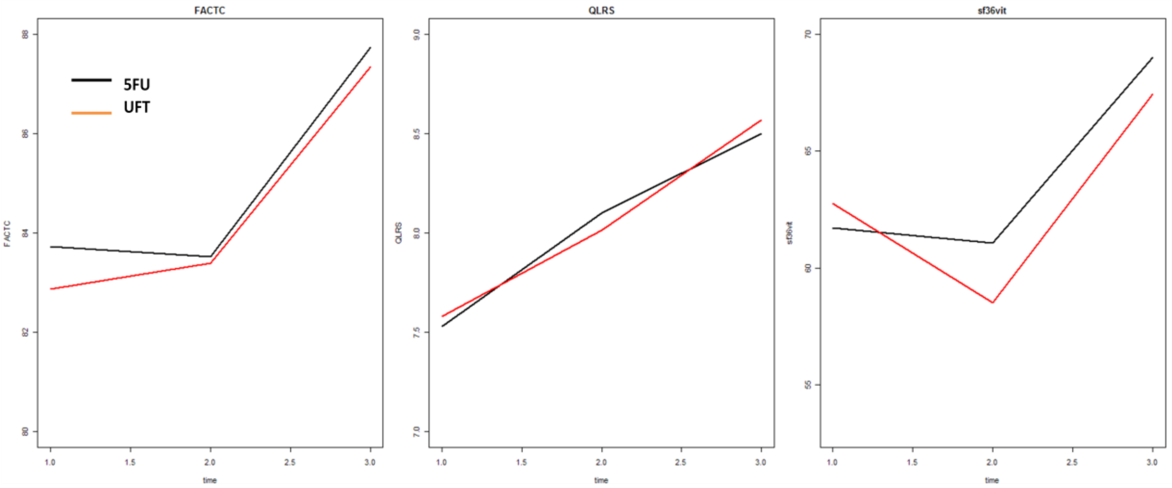

Figure.3 Mean values of the FACTC, QLRS and sf36-vitality scores in the incomplete data.

From model-based multiple imputation 20 datasets were obtained, 6 subscales of FACT-C scores were added and multiply 100/140 to normalize to FACTC_n score with 0-100 scale. The mean values of FACTC_n score, QLRS and sf36 vitality score summarized in **Figure 4,** and **Table 2.**

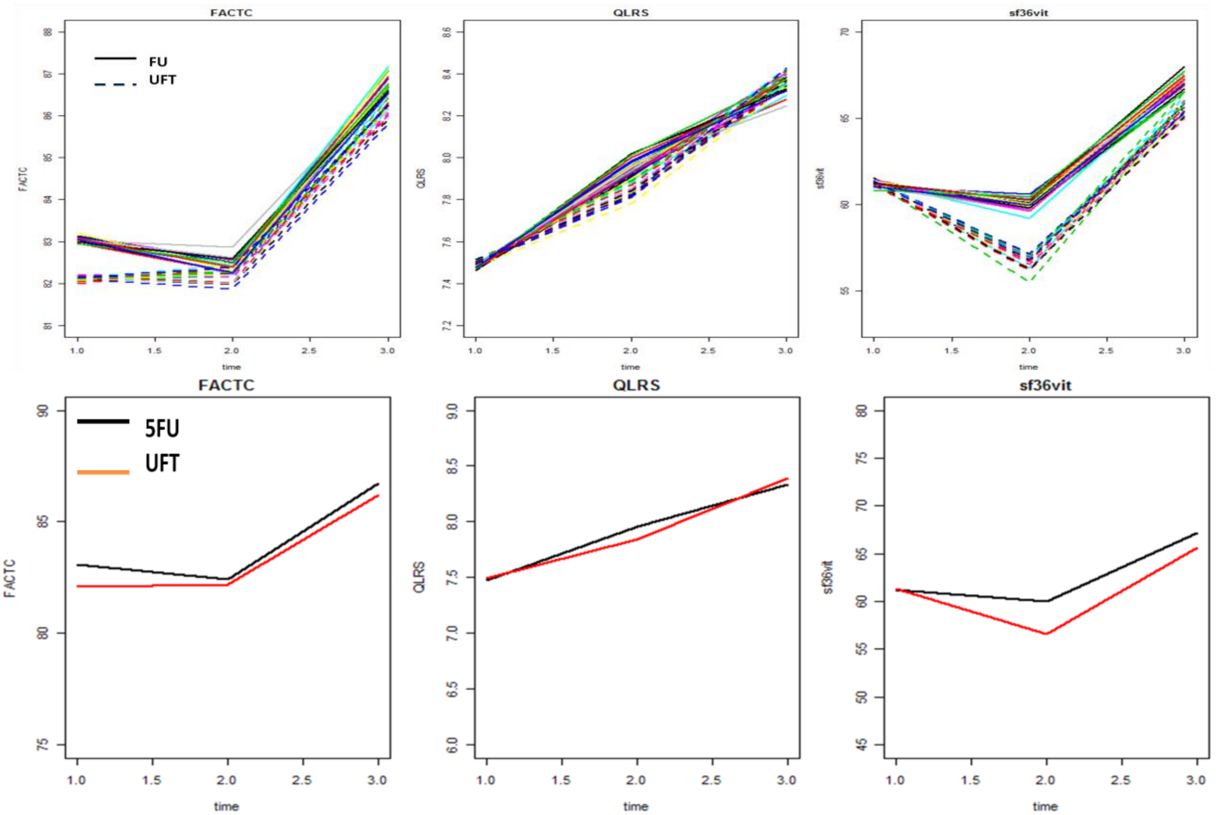

Figure 4. The mean values of the FACTC, QLRS and sf36-vitality scores in the model-based imputed datasets. On the top panels, the mean values from 20 imputed datasets; bottom shows the average to the means.

Table 2. The summary of datasets from model-based multiple imputation

	FACTC_n_score			QLRS			Sf36 Vitality		
	Baseline	chemo	1-year	baseline	chemo	1-year	baseline	chemo	1-year
FU.mean	83.06	82.43	86.74	7.47	7.95	8.34	61.17	60.04	67.21
FU.sd	11.25	12.10	11.78	1.96	1.97	2.03	22.46	22.67	23.35
UFT.mean	82.10	82.19	86.20	7.49	7.84	8.39	61.30	56.62	65.58
UFT.sd	11.29	12.47	11.28	1.81	1.89	1.89	21.97	23.322	21.60

With similar procedure, we summarized the 20 datasets from hot deck imputation in **Figure 5,** and **Table 3.**

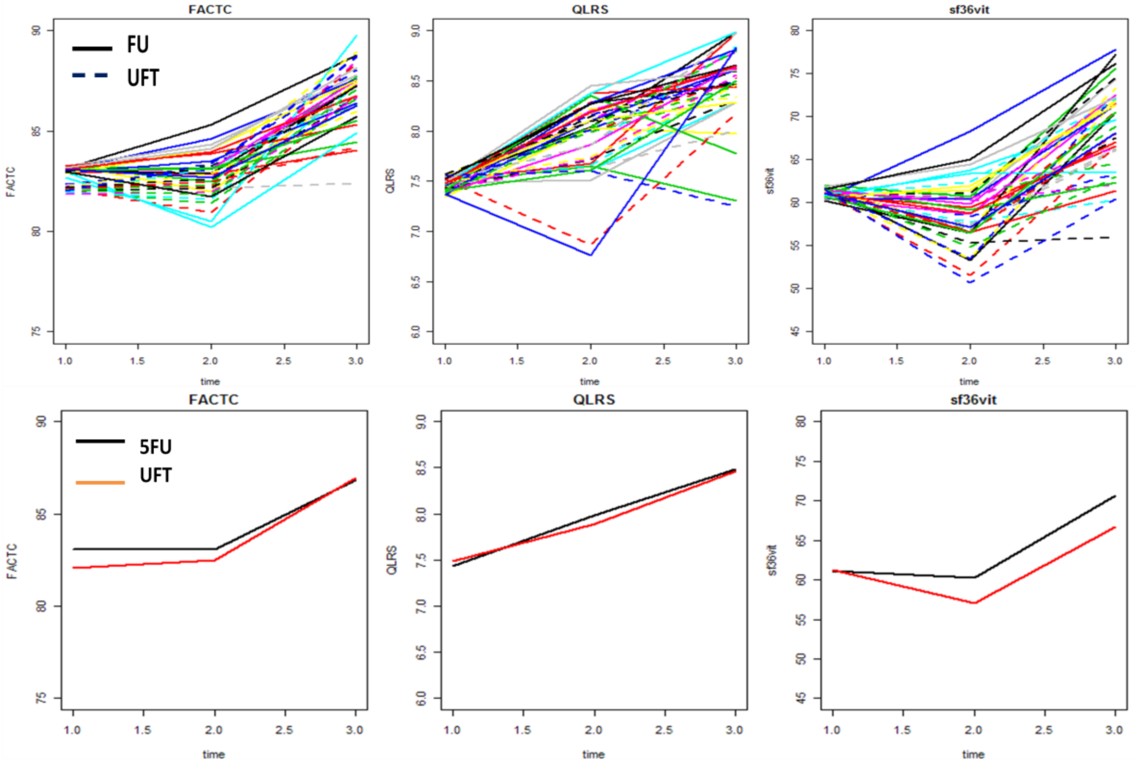

Figure 5. The mean values of the FACTC, QLRS and sf36-vitality scores in the *hot deck* imputed datasets. On the top panels, the mean values from 20 imputed datasets; bottom shows the average to the means.

Table 3. The summary of datasets from hot-deck multiple imputation

	FACTC_n_score			QLRS			Sf36 Vitality		
	baseline	chemo	1-year	baseline	chemo	1-year	baseline	chemo	1-year
FU.mean	83.07	83.08	86.81	7.44	7.98	8.49	61.08	60.29	70.60
FU.sd	11.08	10.64	9.78	1.97	1.83	1.75	22.39	21.18	20.41
UFT.mean	82.08	82.47	86.94	7.49	7.89	8.46	61.21	57.00	66.74
UFT.sd	11.14	11.32	9.65	1.81	1.85	1.63	21.85	23.18	20.79

We noticed that mean values from imputed datasets have more variability, especially in the hot-deck imputations. But when we pooled the mean values from 20 datasets together, they are very similar to the mean values from the original incomplete dataset. The pooled within-imputation standard errors are also very similar. This implies that there might be little difference in health related quality of life between the two treatments arms.

4.3 COMPARE TREATMENT EFFECT WITH LINEAR REGRESSION MODELS

Using the scores (FACTC_nscore, QLRS, and sf36_vitality) at the last time point (1-year) as outcome, the baseline scores and treatment as predictors, linear regression model was fitted to the incomplete data, the results are shown in **Table 4.**

Table 4. The linear regression model for the incomplete data

	FACTC_n score (n=981, R^2=0.1831)				QLRS (n=981, R^2=0.067)				Sf36 Vitality (n=981, R^2=0.172)			
	Est.	S. E.	t	Pr(>|t|)	Est.	S. E.	t	Pr(>|t|)	Est.	S. E.	t	Pr(>|t|)
Intercept	49.10	2.61	18.83	0.000	6.51	0.245	26.61	0.000	43.08	1.999	21.66	0.000
trt	0.099	0.649	0.153	0.878	0.083	0.111	0.752	0.452	-1.960	1.246	-1.573	0.116
baseline	0.454	0.031	14.79	0.000	0.255	0.031	8.310	0.000	0.410	0.029	14.21	0.000

Noticed that there are only 981 observations are used in the models. These are the patients who have filled all the questionnaire at all three time points, 413 in FU treatment arm and 468 in UFT arm (Section **3.1.1, Figure 1**). The model showed patients in UFT arm have 0.1 unit increases in their FACTC score, 0.08 increases in QLRS score, and 1.96 decreases in Sf36 vitality score comparing to FU arm. But all these changes are not statistically significant, with the p-values 0.878, 0.452 and 0.116 respectively. Also we noticed that the R^2 values are very small, which are 0.1831, 0.067, and 0.172 respectively. This means that the linear regression models do not fit the data well. But before we switch to another kind of model, we first check the linear regression models fit the imputed datasets. Models were fitted to 40 imputed datasets from imputation, among which 20 for model-based imputation datasets, 20 for hot deck imputation. Treatment effects in most of the linear models are not significant for FACTC and QLRS in the model-based imputed datasets (**Figure 6**).

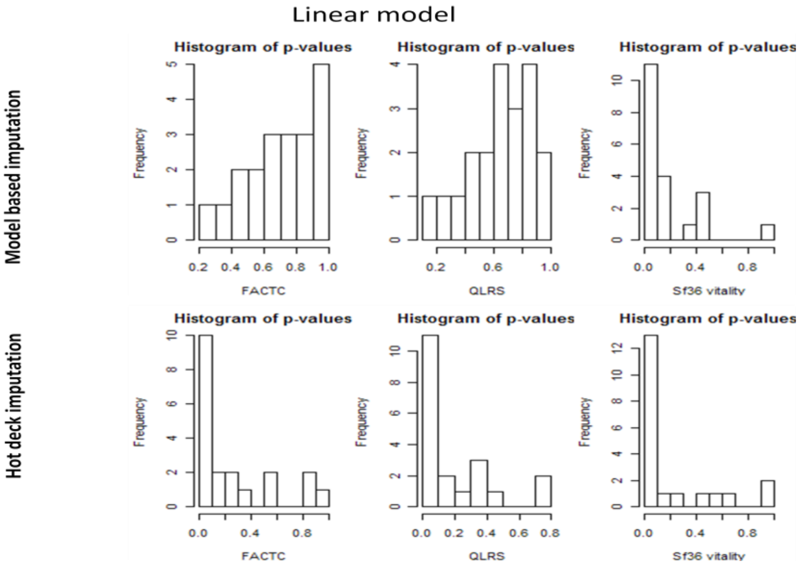

Figure 6. The histograms of p-values of linear models fitting to the imputed data

However there are some models showing that it is significant in FACTC and QLRS in the hot deck imputation datasets. And for Sf36 vitality, more than 11 models showed that treatment effect was close to statistical significance. But the R^2 values were very small, which were just similar to that from model of incomplete data. There was obvious variability between the imputed datasets, we pooled the regression model coefficients information and summarized in **Figure 7** and **Table 5.**

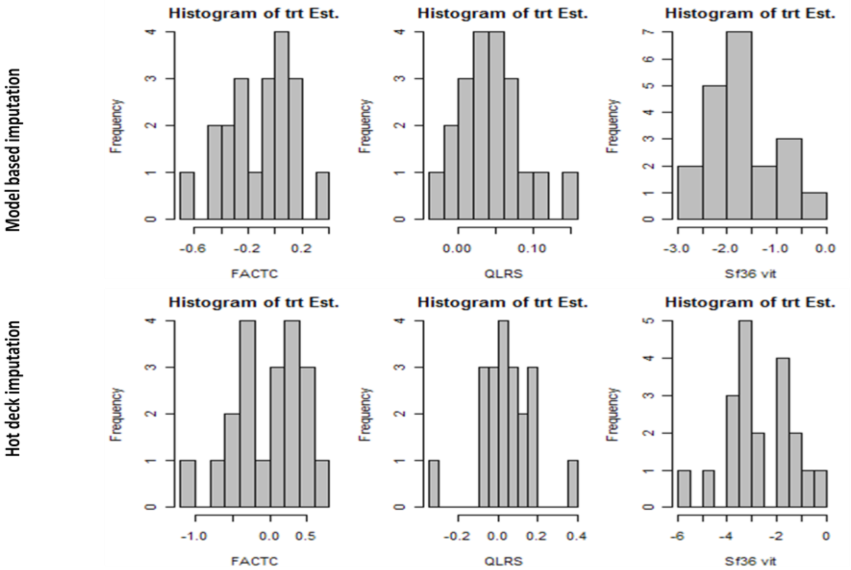

Figure 7. The histograms of estimation of treatment coefficients in linear models fitting to the imputed data

In **Figure 7**, the estimated coefficients β's are basically normally distributed, to estimate if it is significantly different from 0, we need to know its variability. The variability includes two parts, within the imputation and between the imputations. The later part can also be considered variability due to data missing. The variability for both model-based and hot deck imputed datasets were summarized in **Table 5.**

Table 5. The variability of estimation in linear models for the imputed datasets

	Model-based imputation			hot deck imputation		
	FACTC	QLRS	Sf36	FACTC	QLRS	Sf36
$\bar{\beta}$	-0.0999	0.0456	-1.6826	-0.0226	0.0476	-2.7102
$\bar{W}_{20}$	0.2787	0.0094	1.1119	0.3929	0.0116	1.516
B_{20}	0.0634	0.0019	0.4863	0.2079	0.0174	2.0413
T_{20}	0.3389	0.0112	1.5739	0.5905	0.0281	3.4553
γ_{20}	0.1963	0.1769	0.3244	0.3698	0.6508	0.6203

In **Table 5,** $\bar{\beta}$ is the average treatment coefficients from 20 imputed datasets, $\bar{W}20$ is the within imputation variance of treatment effect, B_{20} is between imputation variance, T_{20} is total variability associated with the imputation of the missing values and γ_{20} is the ratio of variability due to missing to that due to the imputation (Section **3.2**). Noticed that γ_{20} in hot deck imputation is almost twice higher than that in the model-based imputation.

Although the treatment effect showed in some imputed datasets is significant, it might just because of the variability of the missing values. To have a consistence inference, we must pool all the information from all 20 datasets. The pooled 95% confidence intervals were calculated (T**able 6**) by:

$$95\% \text{ C.I.} = \bar{\beta} \pm 1.96 *\text{sqrt}(T_{20})$$

Table 6. **The 95% CI of treatment effect from linear models for incomplete and imputed datasets**

	FACTC		QLRS		SF36 vitality	
incomplete	(-1.48,	1.678)	(-0.57,	0.736)	(-4.164,	0.244)
mice.impute	(-1.24,	1.042)	(-0.162,	0.253)	(-4.142,	0.776)
hotdeck.impute	(-1.529,	1.484)	(-0.281,	0.376)	(-6.353,	0.933)

The treatment effects for FACTC, QLRS and Sf36 vitality scores are not significantly different with pooled 95% confidence intervals (-1.241, 1.041), (-0.162, 0.253) and (-4.141, 0.776) respectively for model based- imputed data; and (-1.529, 1.483), (-0.281, 0.376) and (-6.353, 0.933) respectively for hot deck imputation data. This is consistent with the inference from incomplete data and the conclusion from the HRQL study in NSABP C-06 clinical trial (Kopec, 2006). Also the confidence intervals (CIs) for FACTC and QLRS from mice imputed data are narrower than that from incomplete Datasets. The CIs for FACTC and QLRS from hot deck imputed data are similar with incomplete data. But for Sf36, the CIs from imputed dataset did not improve in hot deck imputation.

4.4 COMPARE TREATMENT EFFECT WITH GEE MODELS

GEE (Generalized estimation equations) takes into account the correlation between repeated measures, and can make use of all available observed values, even when there is missing values in that observation. In our data we use two times (during chemotherapy and 1-year after follow up) as repeated measures, unstructured correlation structure is selected and robust variance estimator is applied. The modeling result from the incomplete data is show in **Table 7.**

Table 7. The GEE model for the imputed datasets

Variables	Trt (β)	S.E.	z	P>|z|	95% CI	
FACT_C_n	-0.5475	0.5101	-1.07	0.283	-1.547	0.452
QLRS	0.0073	0.0759	0.10	0.923	-0.141	0.156
36SF-vit	-1.1536	0.9543	-1.21	0.227	-3.024	0.716

Note: No. of obs = 3686, No. of groups = 1527

In **Table 6**, the GEE model utilized all 3686 observations in 1527 patients. Based on this model, patients in UFT arm have 0.55 unit decreases in their FACTC score, 0.007 increases in QLRS score, and 1.15 decreases in Sf36 vitality score comparing to FU arm. But all these changes are not statistically significant, with the p-values 0.283, 0.923 and 0.2276 respectively. The treatment effects were not statistically significant.

GEE models were fitted to each of the imputed datasets. The p-values from each model are summarized in **Figure 8.** But as mentioned above, it is not important for making inference.

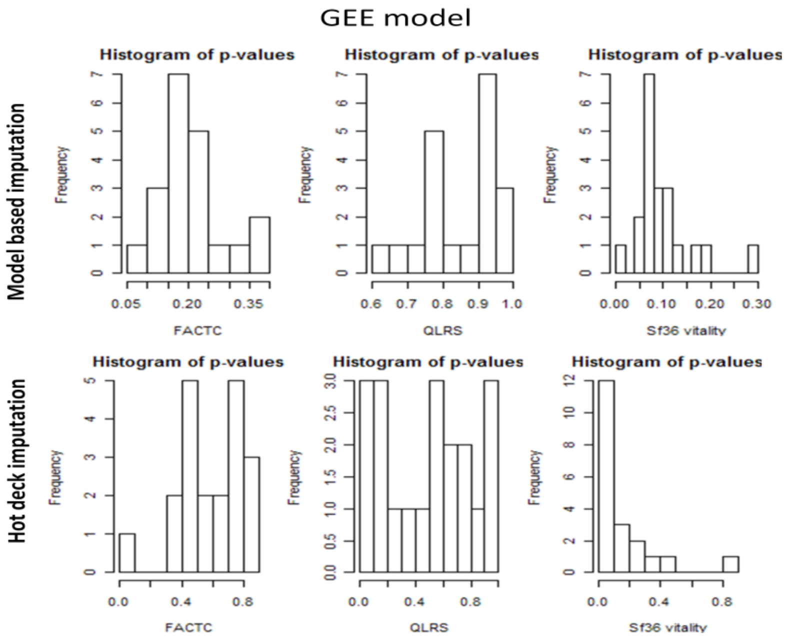

Figure 8. The histograms of p-values of GEE models fitting to the imputed data

The estimated coefficients from GEE models were summarized in **Figure 9.** The treatment effect coefficients variability is summarized in **Table 7.**

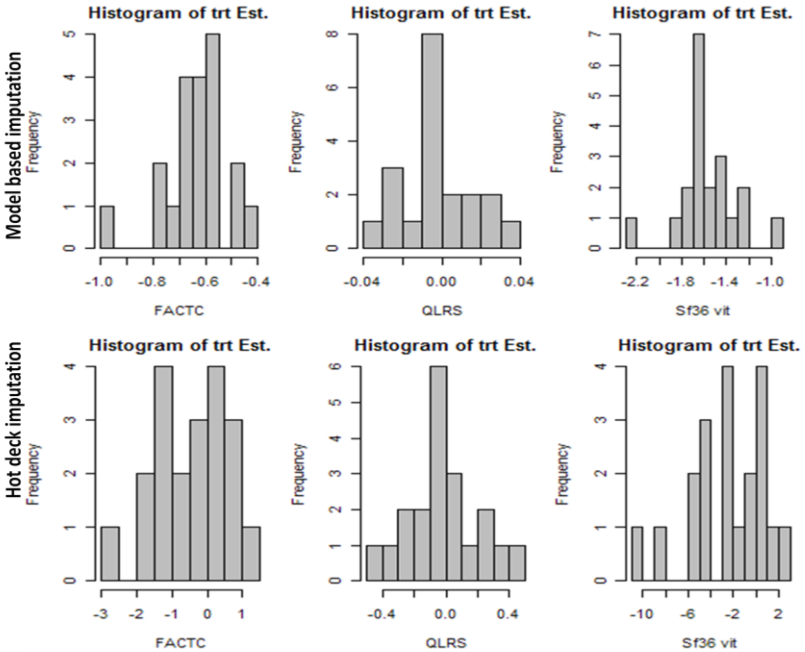

Figure 9. The histograms of estimation of treatment coefficients in GEE models fitting to the imputed data

Table 8. The variability of estimation in GEE models for the imputed datasets

	Model-based imputation			hot deck imputation		
		QLRS	Sf36	FACTC	QLRS	Sf36
$\bar{\beta}$	-0.6346	-0.0023	-1.5765	-0.4406	-0.0182	-2.474
$\bar{W}_{20}$	0.2418	0.0056	0.8346	0.1673	0.0042	0.6458
B_{20}	0.0146	3.00E-04	0.0729	1.0802	0.0502	10.939
T_{20}	0.2557	0.0059	0.9038	1.1935	0.0519	11.038
γ_{20}	0.0601	0.0574	0.0847	0.9051*	0.9672*	0.9910*

Note: *, indicates these values were calculated as B_{20}/T_{20}

Table 9. The 95% CI of treatment effect from GEE models for incomplete and imputed datasets

	FACTC		QLRS		SF36 vitality	
incomplete	(-1.947,	0.852)	(-0.533,	0.547)	(-3.068,	0.762)
mice	(-1.626,	0.357)	(-0.153,	0.148)	(-3.44,	0.287)
hotdeck	(-2.582,	1.701)	(-0.465,	0.428)	(-8.986,	4.038)

The 95% confidence intervals for treatment effect on FACTC, QLRS and Sf36 vitality scores from above with pooled information were calculated: (-1.626, 0.357), (-0.153, 0.148) and (-3.440 0.287) respectively for model based- imputed data; and (-2.582, 1.701), (-0.465, 0.428) and (-8.986, 4.038) respectively, for hot deck imputation data. These CIs showed that the treatments were not significantly different in the quality of life. This is also consistent with the inference from incomplete data and the conclusion from the HRQL study in NSABP C-06 clinical trial (Kopec, 2007).

Note that CI's for FACTC, QLRS and Sf36 vitality from model-based imputed data are narrower than that from incomplete data. The CI's for QLRS from hot deck imputed data are narrower than that from incomplete data. But for FACT and Sf36, the CI's from imputed dataset have not improved or even worsen in hot deck imputation.

5.0 DISCUSSION

5.1 MISSING MECHANISM AND IMPUTATION

5.1.1 Missing mechanisms in HRQL data

Both the original paper and this thesis, concluded that the missing data might not be completely at random (Kopec, 2006). The original paper made the inference based on the response rates in two treatment arms. They authors counted the number of patients response at each time in both arms, using chi-square test to check if the rates are statistically different. The underlying assumption of the test is that all the patients responded at later time responded earlier. In another word, the missing is completely monotonic. But the fact is there are about 10 percent of incompleteness does not belong to monotonic missing pattern (section **4.1.1, 4.1.2**), only 681 out of 766 in FU arm (89%) and 693 out of 761 in UFT arm (90.1%) belong to monotonic pattern. So the test in the original paper is not validated. In this thesis, we use the fisher exact test to test drop-out rate in the monotonic pattern data (129 out of 681 in FU arm vs 94 out of 693 in UFT during chemotherapy) gave a p-value 0.025. This is statistically significant. Based on this test result, the missing patterns and differences in observed scores between complete and drop-out subjects in two arms. Therefore we reached the same conclusion that the data were not completely at random. This supports the conclusion drawn in the published paper.

5.1.2 Imputation missing data

Two multiple imputation approaches were used in this thesis, explicit model-based and implicit model-based (hot deck). In the model-based approach, we keep the fewer level covariates as categorical, and use *linear discriminant analysis (lda)* method to impute the missing values, and for covariates with more than 20 levels, we use *Bayesian linear regression* method. For curious, we also tried to use *Bayesian linear regression (norm*) for all covariates. Surprisingly, the result is very similar (**Figure 10**). The difference is , the imputed values are not integers.

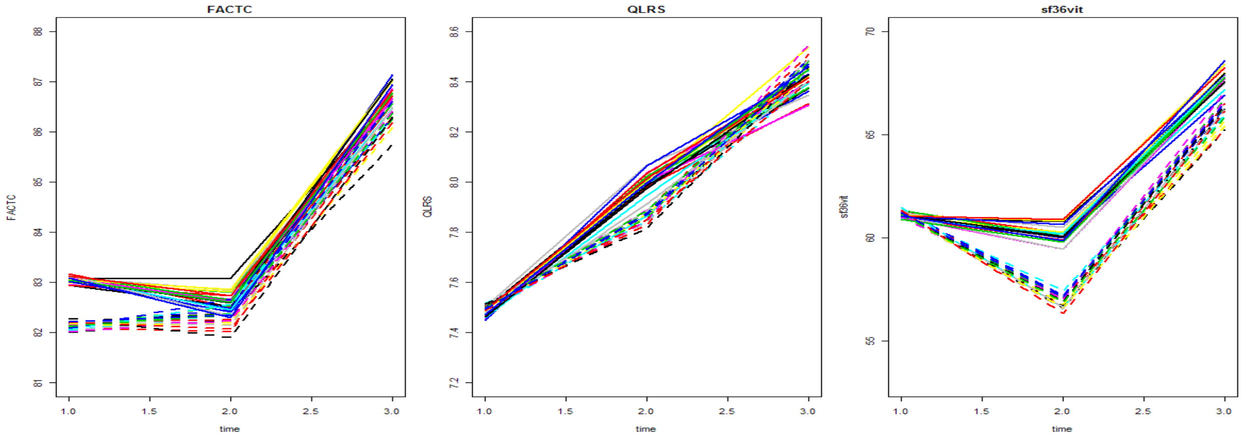

Figure 10. The mean values of the FACTC, QLRS and sf36-vitality scores in the *mice (norm)* imputed datasets. For covariates, the Bayesian linear regression methods were used to impute missing values.

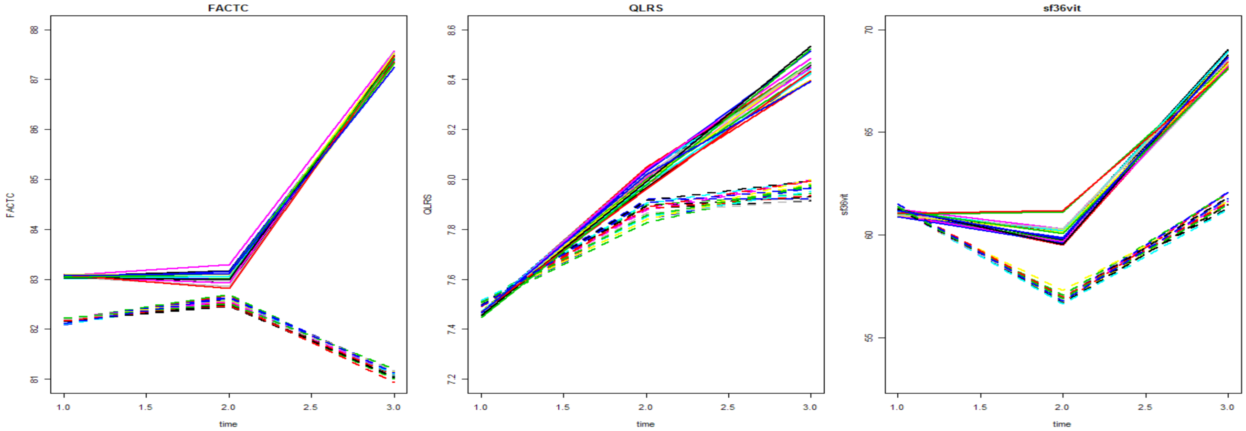

Figure 11. The mean values of the FACTC, QLRS and sf36-vitality scores in the hot deck (un-weighted) imputed datasets. The distance metric is defined with original scales of each covariate.

In hot deck approach, the metric to calculate neighborhood is the most important step. If the metric is not appropriate, it leads to severe bias in the estimation. At first, we tried to used difference of two units in the original scale of each covariates as distance, the mean value from the imputed data are very different from the original data (**Figure 11**), especially at the last time point, where the missing rate is much higher. For FACT-C score, since it is calculated from 6 subscales scores, the bias is cumulative, so it appears the largest bias. But we also notice that the 20 imputed datasets are more consistent than that from weighted hot deck imputation (Section **4.2**, **Figure 5**). It is possible to find a definition of distance, with which we can impute the missing data unbiased and consistently.

Compare the model-based imputation and hot deck imputation data, we noticed that the model based imputation is more consistent in mean values of the covariates, and estimation of coefficients. Hot deck imputation in this thesis, has more variability. But it is difficult to say which one is better when we only look at the mean of the variables. Because the missing value is unknown, it is possible that it changes from the minimal to the maximal possible value. The imputation approaches are just approximately estimation based on observed values.

5.2 ESTIMATION OF TREATMENT EFFECT

The estimation of treatment effect from both the original incomplete data and imputed datasets is quite consistent. The treatment effect on the HRQL is not significantly different in two arms. People may argue that it is not necessary to do a multiple imputation. But it is worthwhile to because the multiple imputation approaches not only consistently estimate the treatment effect, it also successfully account for the uncertainty due the missing. It utilizes the observed data at the same occasion. In the GEE model with model-based imputed datasets, less than 10 percent of the variability is due to the imputation. While for hot deck imputed datasets, more than 95% variability is due to the imputation. But the inferences are the same.

The 95% confidence intervals also indicate in both model based and hot deck imputation, Sf36 estimation efficiency has not improved. It may because that the scale of Sf36 vitality scores is so different from other covariates that the model used in the imputation process cannot account for it. For example, in the hot deck imputation used in this thesis, the distance definition in neighborhood calculation down-weighted the Sf-36 scores. It may have a balanced point that we can improve the estimation by modifying the definition of metric to account for both Sf36 vitality and other covariates.

6.0 CONCLUSION

Based on Missing at random (MAR) mechanism, the missing data in HRQL study in the NSABP C-06 did not introduce significant bias since the estimates of the treatment effect for health related quality of life (HRQL) based on imputed datasets were similar to those based on the original incomplete dataset. However, the model-based multiple imputation provided estimates of confidence intervals that are narrower than that from incomplete data, which means the estimation efficiency is improved through multiple imputation.. In the GEE estimation with model-based imputed datasets, less than 10 percent of the variability is due to the imputation. While for hot deck imputed datasets, more than 95% variability is due to the imputation. But the inferences on the treatment effect are the same. So the two imputation approaches successfully account for the uncertainty due to the incompleteness Therefore, the multiple imputation, especially the model based imputation, was worthwhile since it gave more reasonable and efficient estimates for the treatment effect.

Similar with that from inference of survival outcomes, we didn’t find significant difference in treatment effect on the HRQL. So we could not make decision on treatment selection based on HRQL study. Further study (such as convenience of care study) is needed for the treatment selection.

APPENDIX A

The estimation from imputed datasets

Table A.1 Linear regression result for 20 datasets from model-based imputation

impute No.	FACTC_n score				QLRS				Sf36 Vitality			
	Est.	S.E.	t	P>|t|	Est.	S.E.	t	P>|t|	Est.	S.E.	t	P>|t|
1	-0.23039	0.527373	-0.44	0.662	-0.02386	0.096816	-0.25	0.805	-1.85908	1.052378	-1.77	0.078
2	0.151588	0.544743	0.28	0.781	0.046598	0.102158	0.46	0.648	-1.51574	1.061997	-1.43	0.154
3	-0.21998	0.519336	-0.42	0.672	0.05164	0.095435	0.54	0.589	-0.6978	1.03438	-0.67	0.5
4	-0.03723	0.544966	-0.07	0.946	0.078804	0.094776	0.83	0.406	-1.69423	1.082734	-1.56	0.118
5	0.035577	0.534117	0.07	0.947	0.105635	0.096436	1.1	0.274	-0.86178	1.053863	-0.82	0.414
6	-0.04444	0.529144	-0.08	0.933	-0.01251	0.093998	-0.13	0.894	-2.0957	1.054638	-1.99	0.047
7	-0.40112	0.52234	-0.77	0.443	-0.0067	0.098187	-0.07	0.946	-1.99244	1.055163	-1.89	0.059
8	-0.21547	0.521455	-0.41	0.68	0.010115	0.091878	0.11	0.912	-2.68003	1.038695	-2.58	0.01
9	0.175258	0.535395	0.33	0.743	0.012764	0.098253	0.13	0.897	-0.80396	1.044164	-0.77	0.441
10	0.03104	0.525341	0.06	0.953	0.049433	0.098153	0.5	0.615	-1.05863	1.046653	-1.01	0.312
11	0.399202	0.515747	0.77	0.439	0.061116	0.09491	0.64	0.52	-0.06305	1.021956	-0.06	0.951
12	-0.61091	0.522109	-1.17	0.242	0.036913	0.0981	0.38	0.707	-2.07045	1.030552	-2.01	0.045
13	-0.31595	0.519906	-0.61	0.543	0.07304	0.096739	0.76	0.45	-1.9103	1.059375	-1.8	0.072
14	-0.33802	0.530672	-0.64	0.524	0.048453	0.099774	0.49	0.627	-2.26283	1.060177	-2.13	0.033
15	-0.03673	0.5379	-0.07	0.946	0.039569	0.095727	0.41	0.679	-2.11443	1.095749	-1.93	0.054
16	0.091398	0.54295	0.17	0.866	0.158516	0.099597	1.59	0.112	-1.414	1.096578	-1.29	0.197
17	0.079644	0.522924	0.15	0.879	0.031592	0.097327	0.32	0.746	-2.86995	1.015225	-2.83	0.005
18	-0.49076	0.532054	-0.92	0.356	0.033714	0.093809	0.36	0.719	-1.80624	1.077008	-1.68	0.094
19	-0.13558	0.508688	-0.27	0.79	0.018429	0.096152	0.19	0.848	-2.20064	1.038961	-2.12	0.034
20	0.115224	0.519522	0.22	0.825	0.09818	0.097629	1.01	0.315	-1.67989	1.064647	-1.58	0.115

Table A.2 Linear regression result for 20 datasets from hot deck imputation

	FACTC_n score				QLRS				SF-36 Vitality			
Imputation No.	Est.	S. E.	t	Pr(>|t|)	Est.	S. E.	t value	Pr(>|t|)	Estimate	S. E.r	t value	Pr(>|t|)
1	-0.49011	0.626896	-0.78181	0.434502	-0.03222	0.107402	-0.29996	0.764263	-2.60508	1.236512	-2.1068	0.035371
2	0.49393	0.624859	0.790466	0.429433	-0.05003	0.106423	-0.47008	0.638394	-1.17572	1.211565	-0.97042	0.332061
3	-0.23774	0.622577	-0.38186	0.702642	0.363166	0.125732	2.888422	0.003951	-3.44648	1.248822	-2.75978	0.005884
4	0.100873	0.632678	0.159439	0.873354	0.007181	0.10515	0.06829	0.945568	-4.84079	1.241784	-3.89825	0.000103
5	-0.63632	0.637048	-0.99885	0.318095	0.151464	0.105451	1.436342	0.151202	-3.63157	1.222905	-2.96962	0.003049
6	0.123642	0.630714	0.196035	0.84462	0.063545	0.106031	0.59931	0.549095	-3.89065	1.200696	-3.24033	0.001231
7	0.110447	0.641968	0.172044	0.863436	0.156958	0.107304	1.462747	0.143835	-1.64901	1.225817	-1.34524	0.178839
8	-1.14425	0.622268	-1.83884	0.06622	-0.0016	0.10498	-0.01526	0.987825	-3.44029	1.272209	-2.70418	0.006958
9	-0.23873	0.622282	-0.38363	0.701328	0.019314	0.104935	0.18406	0.854001	-2.57707	1.304355	-1.97574	0.048445
10	0.634229	0.629677	1.007229	0.314056	0.072876	0.104607	0.696658	0.486171	-0.94433	1.197251	-0.78875	0.430437
11	0.451862	0.616307	0.733176	0.463614	0.18986	0.105419	1.801009	0.071987	-1.31777	1.185746	-1.11134	0.266673
12	0.471216	0.629258	0.748843	0.454119	-0.00503	0.106016	-0.04745	0.962165	-1.78477	1.247104	-1.43113	0.152689
13	0.201571	0.630528	0.319687	0.749269	-0.056	0.106182	-0.52741	0.598023	-0.25131	1.221085	-0.20581	0.836978
14	-0.17659	0.621445	-0.28416	0.776343	0.049929	0.105133	0.474909	0.63495	-3.08967	1.206638	-2.56056	0.010589
15	-0.32679	0.636646	-0.5133	0.607847	0.08945	0.104438	0.856485	0.391924	-1.64668	1.258457	-1.3085	0.190991
16	-0.28277	0.626315	-0.45149	0.651731	0.139068	0.106961	1.30018	0.193824	-3.13181	1.203593	-2.60205	0.009397
17	0.381491	0.621094	0.614224	0.5392	0.027971	0.104793	0.266915	0.789587	-5.97738	1.279462	-4.67179	3.37E-06
18	-0.45419	0.615468	-0.73796	0.460703	-0.05842	0.107231	-0.5448	0.586006	-1.50592	1.17976	-1.27646	0.202072
19	0.204384	0.6247	0.327172	0.743603	0.13108	0.10602	1.236372	0.216596	-3.40527	1.201476	-2.83424	0.004681
20	0.362021	0.623081	0.581017	0.561353	-0.30563	0.118037	-2.58923	0.009752	-3.89277	1.270879	-3.06305	0.002247

Table A.3 GEE result for 20 datasets from model-based imputation

	FACTC_n				QLRS				SF36_vit			
impute No.	Est..	Std.	z	P>|z|	Est.	Std.	z	P>|z|	Est.	Std.	z	P>|z|
1	-0.64639	0.491586	-1.31	0.189	-0.00646	0.073665	-0.09	0.93	-1.66014	0.909442	-1.83	0.068
2	-0.45517	0.498241	-0.91	0.361	-0.00486	0.076189	-0.06	0.949	-1.59081	0.912295	-1.74	0.081
3	-0.72082	0.489269	-1.47	0.141	-0.00456	0.075869	-0.06	0.952	-1.39553	0.91331	-1.53	0.127
4	-0.5679	0.492836	-1.15	0.249	-0.0037	0.075518	-0.05	0.961	-1.52692	0.916086	-1.67	0.096
5	-0.59704	0.491214	-1.22	0.224	0.027574	0.075159	0.37	0.714	-1.2194	0.911586	-1.34	0.181
6	-0.69047	0.492797	-1.4	0.161	-0.02277	0.074807	-0.3	0.761	-1.68397	0.918955	-1.83	0.067
7	-0.68611	0.485693	-1.41	0.158	-0.03267	0.075782	-0.43	0.666	-1.6848	0.915407	-1.84	0.066
8	-0.76995	0.490068	-1.57	0.116	-0.00734	0.074948	-0.1	0.922	-1.63537	0.914932	-1.79	0.074
9	-0.57016	0.49084	-1.16	0.245	-0.02066	0.075741	-0.27	0.785	-1.24619	0.91163	-1.37	0.172
10	-0.60185	0.49523	-1.22	0.224	-0.02214	0.074914	-0.3	0.768	-1.45462	0.905892	-1.61	0.108
11	-0.45975	0.491005	-0.94	0.349	0.016807	0.074591	0.23	0.822	-0.96116	0.90961	-1.06	0.291
12	-0.95628	0.494364	-1.93	0.053	-0.00864	0.076024	-0.11	0.909	-1.61648	0.894972	-1.81	0.071
13	-0.64059	0.492839	-1.3	0.194	0.021095	0.075811	0.28	0.781	-1.47641	0.910476	-1.62	0.105
14	-0.65503	0.490479	-1.34	0.182	0.00098	0.075377	0.01	0.99	-1.71844	0.918106	-1.87	0.061
15	-0.68233	0.500986	-1.36	0.173	-0.00656	0.075466	-0.09	0.931	-1.85485	0.926352	-2	0.045
16	-0.64299	0.489176	-1.31	0.189	0.037981	0.075046	0.51	0.613	-1.67353	0.918922	-1.82	0.069
17	-0.55265	0.493061	-1.12	0.262	-0.01907	0.074298	-0.26	0.797	-2.28241	0.90457	-2.52	0.012
18	-0.77105	0.490776	-1.57	0.116	0.01082	0.074702	0.14	0.885	-1.79929	0.912626	-1.97	0.049
19	-0.58821	0.487493	-1.21	0.228	-0.00838	0.073989	-0.11	0.91	-1.60757	0.926305	-1.74	0.083
20	-0.43778	0.487245	-0.9	0.369	0.005687	0.074069	0.08	0.939	-1.44308	0.918755	-1.57	0.116

Table A.4 GEE result for 20 datasets from hot deck imputation

	FACTC_n				QLRS				SF36_vit			
impute No.	Coef.	Std.	z	P>|z|	Coef.	Std.	z	P>|z|	Coef.	Std.	z	P>|z|
1	-1.72334	0.406369	-4.24	0	-0.13125	0.060712	-2.16	0.031	0.04474	0.81361	0.05	0.956
2	0.222931	0.414782	0.54	0.591	-0.08837	0.060371	-1.46	0.143	0.968554	0.798574	1.21	0.225
3	-1.13228	0.409532	-2.76	0.006	0.463753	0.075787	6.12	0	-4.18205	0.776813	-5.38	0
4	-0.25421	0.402128	-0.63	0.527	0.018123	0.062179	0.29	0.771	-8.00821	0.80735	-9.92	0
5	-0.85964	0.403608	-2.13	0.033	0.316848	0.061417	5.16	0	-5.41243	0.796378	-6.8	0
6	-0.02738	0.400256	-0.07	0.945	0.060129	0.063267	0.95	0.342	-2.88212	0.774273	-3.72	0
7	0.613598	0.405695	1.51	0.13	0.254958	0.065243	3.91	0	-2.64344	0.786695	-3.36	0.001
8	-2.97839	0.419934	-7.09	0	-0.06136	0.066998	-0.92	0.36	-4.339	0.805695	-5.39	0
9	-0.76322	0.401004	-1.9	0.057	-0.22005	0.062691	-3.51	0	0.620087	0.792806	0.78	0.434
10	0.378462	0.408851	0.93	0.355	0.054788	0.063285	0.87	0.387	0.066004	0.842581	0.08	0.938
11	0.102989	0.411285	0.25	0.802	0.149368	0.06406	2.33	0.02	1.290604	0.810835	1.59	0.111
12	0.509415	0.402403	1.27	0.206	-0.10453	0.062407	-1.67	0.094	-1.53751	0.80758	-1.9	0.057
13	1.277802	0.418179	3.06	0.002	-0.35015	0.063046	-5.55	0	2.785472	0.80215	3.47	0.001
14	-1.06248	0.405885	-2.62	0.009	-0.06517	0.061918	-1.05	0.293	-2.20508	0.798129	-2.76	0.006
15	-1.22624	0.420788	-2.91	0.004	-0.0432	0.062501	-0.69	0.489	-0.46023	0.782079	-0.59	0.556
16	-1.12224	0.403148	-2.78	0.005	-0.01908	0.060901	-0.31	0.754	-4.6979	0.789299	-5.95	0
17	0.554037	0.413869	1.34	0.181	-0.09028	0.061321	-1.47	0.141	-10.4908	0.858381	-12.22	0
18	-1.72349	0.411093	-4.19	0	-0.46902	0.073024	-6.42	0	-0.34229	0.771194	-0.44	0.657
19	-0.09508	0.411256	-0.23	0.817	0.20978	0.065591	3.2	0.001	-2.95689	0.812827	-3.64	0
20	0.496331	0.409049	1.21	0.225	-0.2493	0.076701	-3.25	0.001	-5.10794	0.838805	-6.09	0

APPENDIX B

R-code for nearest neighborhood hot deck imputation

```
##### define function to find the missing pattern
ind.fun<-function(a){              ### find the missing indicator matrix
for(i in 1:nrow(a)){
for(j in 1:27){
if(!is.na(a[i,j])) d[i,j]<-1
}} return(d)
}
###################### define missing pattern#############
p1<- 0
p2<-c(3,6,9,12,15,18,21,24,27)
p3<-c(2,5,8,11,14,17,20,23,26)
p4<-c(2,3,5,6,8,9,11,12,14,15,17,18,20,21,23,24,26,27)
p5<-c(1,4,7,10,13,16,19,22,25)
p6<-c(1,3,4,6,7,9,10,12,13,15,16,18,19,21,22,24,25,27)
p7<-c(1,2,4,5,7,8,10,11,13,14,16,17,19,20,22,23,25,26)
pattern<-list(p1,p2,p3,p4,p5,p6,p7)  ### this pattern list will be used in impute the different missing value in the same subject

####   define function for search nearest neighbor ###############################
searchcloseby<-function(value,set)
     {
          dist<-c(0,0.1,0.2,0.3,0.5, 0.7, 1,2,3,4,5,6,7,8,9,10)        ### set the distance criteria for stop the loop
          number.obs<-27-(length(value[is.na(value)]))                 ### the number of variable used to calculate the distance

          value[c(7:12,16:24)]<-value[c(7:12,16:24)]/2.8               ###set the weight for calculate distance in the vactor to be imputed
          value[25:27]<-          value[25:27]/10
          value[13:15]<-value[13:15]*5/4
          A<-sum(value, na.rm=T)

          set1<-set                                                    ###set the weight for calculate distance in the donor metrix
          set1[,c(7:12,16:24)]<-set1[,c(7:12,16:24)]/2.8
          set1[,25:27]<-set1[,25:27]/10
          set1[,13:15]<-set1[,13:15]*5/4
          B<-rowSums(set1[,pattern[9-d][[1]]])

          distance<-abs(B-A)/number.obs                                #### calculate the normalized distance (in 0.00-10.00 scale,)
          w<-cbind(distance,set)
          w1<-w[order(w[,1]),]   ### order the distance so the nearest subject at the top

          s<-1
          while(nrow(w1[(w1[,1]<=dist[s]),])< 2 ){s<-s+1}
          subset<- w1[(w1[,1]<=dist[s]),]
          subset              ### output
     }
############  define function for hot deck imputation ##############################
hot.deck.impute<-function(dataset,dataout)
{
  combined<-dataset[Ind==a[1],]  ##set the intitial dataset
       for(d in 2:7)                                                   ### pattern 2 to pattern 7 have missing values
       {
          impset<-dataset[Ind==a[d],]                                  ### identify the subset which need to be imputed
          impset[is.na(impset)]<-0                                     ### change NA to 0, so the imputede value can be added on
          mdvector<-rownames(impset)                                   ### identify the subject of the subset
          codeset<-dataset[Ind==a[1],]                                 ### use the complete cases as donor (pattern 1)

          num.cases<-length(mdvector)                                  ### the number of subject in this pattern need to be imputed
```

```
            random.numbers<-matrix(runif(num.cases*27,0,1),nrow=num.cases,ncol=27,)
                    ### generate a matrix of random Number which has the same dimension as the subset to be imputed
            income.impute<-matrix(0,nrow=num.cases,ncol=27, byrow=TRUE,dimnames = list(rownames(impset), colnames(impset)))
                    ### generate a matrix of 0 which has the ame dimension the subset to be imputed

            for (i in 1:num.cases)
            {
                    icid<-mdvector[i]                                ### the subject names in the subject to be imputed
                    impcode<-dataset[rownames(dataset)==icid,] ### identify the vector of the subject

                    subset<-searchcloseby(impcode,codeset)      ### looking for the nearest neighborhood
                    k<-nrow(subset)

                    for(m in pattern[d][[1]])                   ### impute the missing values in the subject through random draw.
                    {                                           ### each missing value is imputed by a independent random draw
                    selected<-ceiling(k*random.numbers[1,m])  ###from the same neighborhood
                    income.impute[(rownames(impset)==icid),m]<-subset[selected,(m+1)]
                    }
            }
            imputed.set<-impset+income.impute                   ### update the imputed subset
            combined<-rbind(combined,imputed.set)               ### combined it to the initial dataset and update the dataset
   }
   dataout<-combined
   dataout
}
########################## imputation FU ####################################
setwd("C:/thesis/before_imputatioin/")
FU <- read.table("5fu_before_impute_3_14.csv", header=TRUE, sep=",")
Fu.new<-FU[,2:28]

d<-matrix(0, nrow=nrow(Fu.new),ncol=27)
ind01<-ind.fun(Fu.new)                       ### the missing pattern of the 5fu data
Ind<-0
for(i in 1:27){
Ind<- Ind+ ind01[,i]*(2**(28-i))}


a=unique(Ind)  ###    will give the unique patterns represented by the indicator
a<-a[order(a, decreasing=T)]

set.seed(1234)
n.iteration<-20
for(p in 1:n.iteration)
{
fu.impute<-hot.deck.impute(Fu.new,dataout)
write.csv(fu.impute, file=paste("fu_weighted_impute_set_", p, ".csv", sep=""))
}
######################  impute uft data  ##################################
setwd("C:/thesis/before_imputatioin/")

UFT <- read.table("uft_before_impute_3_14.csv", header=TRUE, sep=",")
UFT.new<-UFT[,2:28]
d<-matrix(0, nrow=nrow(UFT.new),ncol=27)

ind01<-ind.fun(UFT.new)                      ### the missing pattern of the uft data
Ind<-0
for(i in 1:27){
Ind<- Ind+ ind01[,i]*2**((28-i))}

a<-a[order(a, decreasing=T)]

set.seed(1234)
n.iteration<-20
for(p in 1:n.iteration)
{
uft.impute<-hot.deck.impute(UFT.new,dataout)
write.csv(uft.impute, file=paste("uft_weighted_impute_set_", p, ".csv", sep=""))
}
```

BIBLIOGRAPHY

Jacek A. Kopec, Greg Yothers, Patricia A. Gaanz, et al: Quality of Life in Operable Colon Cancer Patients Receiving Oral Compared With Intravenous Chemotherapy: Results From National Surgical Adjuvant Breast and Bowel Project Trial C-06. *J. Clin. Oncol.* (2007) 25 (4): 424-430

Lembersky BC, Wieand HS, Petrlli NJ, et al: Oral Uracil and Tegafur Plus Leucovorin Compared With Intravenous Fluorouracil and Leucovorin in Stage II and III Carcinoma of the Colon: Results From National Surgical Adjuvant Breast and Bowel Project Protocol C-06. *J. Clin. Oncol.* (2006) 24 (13): 424-430

Tsai-Shen Yang, Jeng-Yi Wang, Reiping Tang, Kuan-Cheng Hsu and Jen-Shi Chen: Oral Uracil/Ftorafur (UFT) Plus Leucovorin as First-line Chemotherapy and Salvage Therapy with Weekly High-dose 5-Fluorouracil/Leucovorin for the Treatment of Metastatic Colorectal Cancer. *Jpn. J. Clin. Oncol.* (2002) 32 (9): 352-357

Wolmark N, Rockette H, Fisher B, et al: The benefit of leucovorin-modulated fluorouracil as postoperative adjuvant therapy for primary colon cancer: results from National Surgical Adjuvant Breast and Bowel Project protocol C-03. *J. Clin. Oncol.* (1993)11: 1879-1887

Wolmark N, Bryant J, Smith R, et al: Adjuvant 5-Fluorouracil and Leucovorin With or Without Interferon Alfa-2a in Colon Carcinoma: National Surgical Adjuvant Breast and Bowel Project Protocol C-05. *J.Natl. Cancer Inst.* (1998)90:1810-1816

Ward WL, Hahn EA, Mo F, et al: Reliability and validity of the Functional Assessment of Cancer Therapy-Colorectal (FACT-C) quality of life instrument. *Qual. Life Res*. (1999) 8 (3): 181-195

Ware JE Jr, Sherbourne CD: The MOS 36 item short-form health survey (SF-36): I, conceptual framework and item selection. *Med. Care* (1992) 30 (6): 473-483

Rubin, DB: Inference and missing data. *Biometrika.* (1976) 63: 581-592

Enders CK, Applied Missing Data Analysis. (2010) New York, London: *The Guilford Press*

Robins JM, Rotnitzky A, Zhao LP: Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data. *(1995) J. American Statistical Association.* 90(429): 106-115

Little JL, Rubin DB: Statistical Analysis with Missing Data, 2^{nd} ed. Hoboken, NJ: *Wiley*