

Impute Incomplete Patient-Level Medical Cost - Assessment of Multiple Imputation Techniques

Shanyue Guan¹, Pascale Peeters², Paola Pedotti¹, Anke van Engen¹

¹ Quintiles Consulting, Siriusdreef 10, 2131 WT Hoofddorp, The Netherlands, ² Quintiles Consulting, 3-5 rue Maurice Ravel, 92594 Levallois-Perret Cedex, France

Abstract

OBJECTIVES: To address the common issue of incomplete data in cost-effective analysis, we imputed missing cost components at the patient level using multiple imputation (MI) techniques.

METHODS: A study cohort with concomitant medication, hospitalisation and outpatient-visit costs was derived from the population of a randomised clinical trial comparing two treatments. On a total of 132 subjects without missing data, a pattern of missingness was created so that 25% of subjects had missing hospitalisation costs and 50% (including the above 25%) had missing concomitant medication costs. The average total costs (sum of the three components) obtained using MI techniques (propensity score [PS], regression and Markov Chain Monte Carlo [MCMC]), complete case analysis (CCA) and available case analysis (ACA) were compared with actual costs. In imputation models, response variables were the logarithm-transformed medication and hospitalisation costs; covariates were age, gender, race, region, treatment arm, number of adverse events, survival time, treatment discontinuation, and outpatient-visit cost (logarithm-transformed).

RESULTS: Average total costs made up of three cost components are given by treatment A versus treatment B. Actual total cost: £16,527 vs. £18,484. Estimated costs: CCA: £17,317 vs. £15,400; ACA: £13,407 vs. £15,361; MI-PS: £16,941 vs. £19,156; MI-Regression: £16,404 vs. £19,056; MI-MCMC: £16,584 vs. £18,947.

CONCLUSIONS: Treatment B was actually more expensive than treatment A. CCA gave opposite results; while ACA underestimated total costs. Regression gave better results than PS, as a regression model was fitted for each missing cost component, with the previous variables as covariates. MCMC using Bayes' theorem with a non-informative prior makes minimum assumption for the data and gave the best imputation results. Under the assumption of missing at random, MCMC could be a useful imputation technique applied to the patient-level missing costs, to permit a more realistic cost-effective analysis.

Data source

- Open, multicenter phase III, parallel group clinical trial
- Subject characteristics and resource utilization costs were selected from the datasets
- 132 subjects had complete 1-year total costs made up of outpatient visit, hospitalisation and concomitant medication costs
- 50% subjects were selected to had one or more missing cost components therefore their total costs were unknown
- Missingness was created in a missing completely at random (MCAR) manner

Objectives

- To estimate the total average cost (£) per treatment group
- To perform a cost-effectiveness analysis

Table 1: Patterns of missing cost components

Patterns	Outpatient visit	hospitalisation	Concomitant medication
Pattern 1	O	O	O
Pattern 2	O	O	M
Pattern 3	O	M	M

*footnote: O-observed; M-missing

Methods

Naive methods

➤ **Available Case Analysis (ACA):** ignores missingness and treats available costs as total costs

➤ **Complete Case Analysis (CCA):** deletes subjects with missing cost components

Multiple imputation techniques

➤ **Regression:** for a continuous variable Y_j with missing values, a regression model is fitted with the observed values for the variable Y_j and its covariates X_1, X_2, \dots, X_k and based on the fitted regression model, a new regression model is simulated from the posterior predictive distribution of the parameters (regression parameter estimates and associated covariance matrix) and is used to impute the missing values for each variable.

➤ **Propensity Score (PS):** for a variable with missing values, a propensity score is generated for each observation to estimate the probability that the observation is missing. The observations are then grouped based on these propensity scores, and an approximate Bayesian bootstrap imputation is applied to each group. It uses only the covariate information associated with whether the imputed variable values are missing and does not use correlations among variables.

➤ **Markov Chain Monte Carlo (MCMC)**

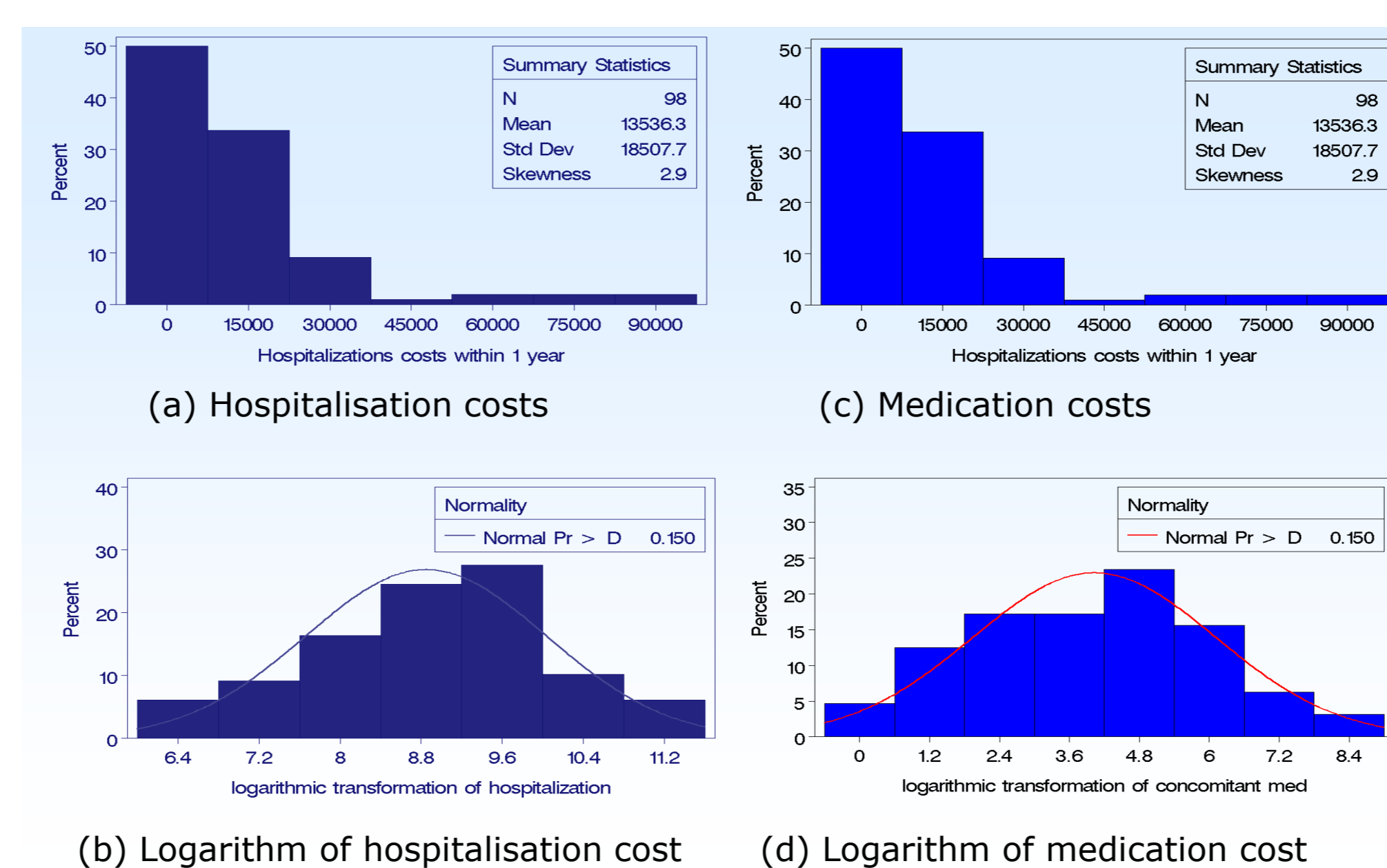
➤ Markov Chain Monte Carlo (MCMC): imputes either all missing values or just enough missing values to make the imputed data set have a monotone missing pattern. By using Bayes' theorem, one can simulate the entire joint posterior distribution of the unknown quantities and obtain simulation-based estimates of posterior parameters that are of interest. Assuming that the data are from a multivariate normal distribution, data augmentation can be applied to Bayesian inference with missing data by repeating the following steps:

- the imputation I-step: given an estimated mean vector and covariance matrix, the I-step simulates the missing values for each observation independently;
- the posterior P-step: given a complete sample, the P-step simulates the posterior population mean vector and covariance matrix. These new estimates are then used in the next I-step.

Procedure

- 10 imputations were applied
- Covariates included in the imputation models were treatment arm, age, number of adverse events, gender, race, geographical area, study terminated, survival duration and outpatient visit cost
- Three cost components were logarithm-transformed to achieve normality assumptions required by the Regression and MCMC methods

Figure 1: Logarithm-transformation of costs



Results

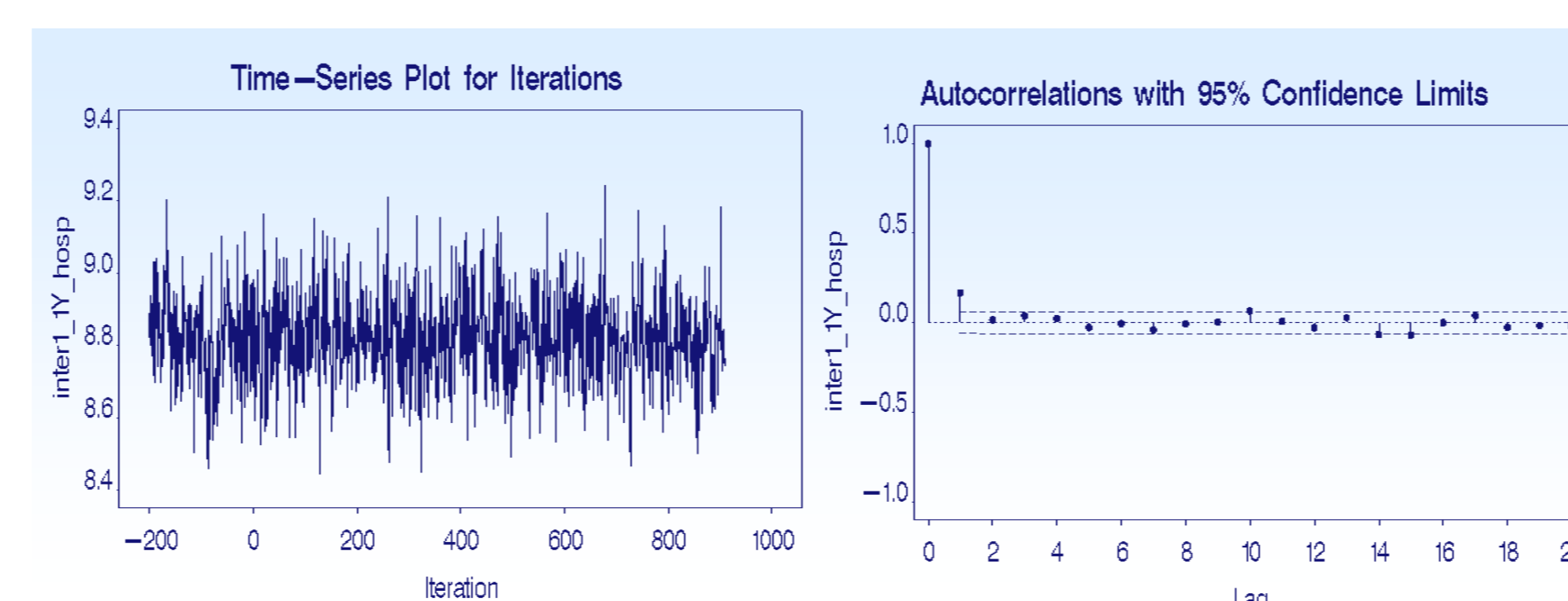
- ACA substantially underestimated the average total costs
- By deleting half of the subjects CCA showed opposite result from the true cost that was treatment B should be more expensive than treatment A
- Regression gave closest results to the true costs than did propensity score
- MCMC using Bayes' theorem with a non-informative prior making minimum assumption for the prior distribution gave the best estimation.

Table 2: Mean total cost estimation by treatment

	Mean cost of group A (£)	Mean cost of group B (£)
True cost	16,527	18,484
ACA	13,407	15,361
CCA	17,317	15,400
Regression	16,404	19,056
Prop. Score	16,941	19,156
MCMC	16,584	18,947

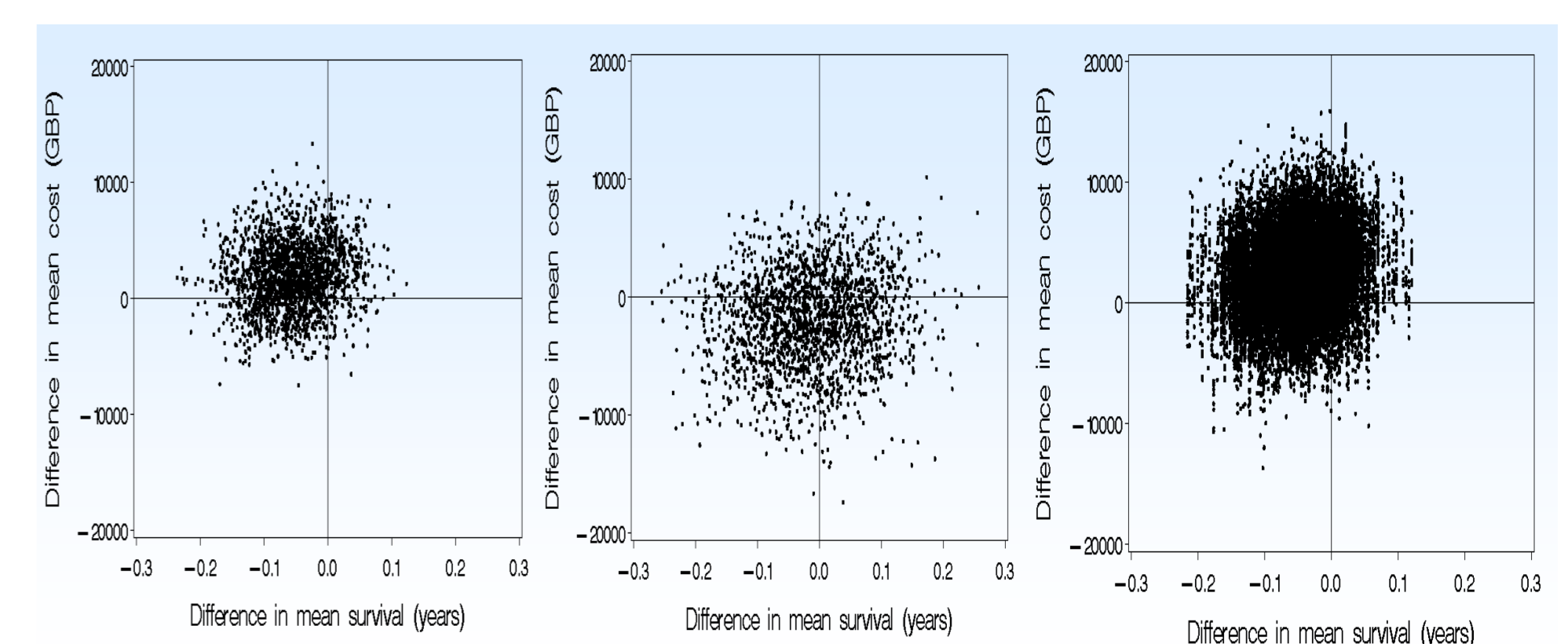
➤ For MCMC, time series plot and autocorrelation showed no correlation between the iterations and convergence was achieved.

Figure 2: Convergence in MCMC



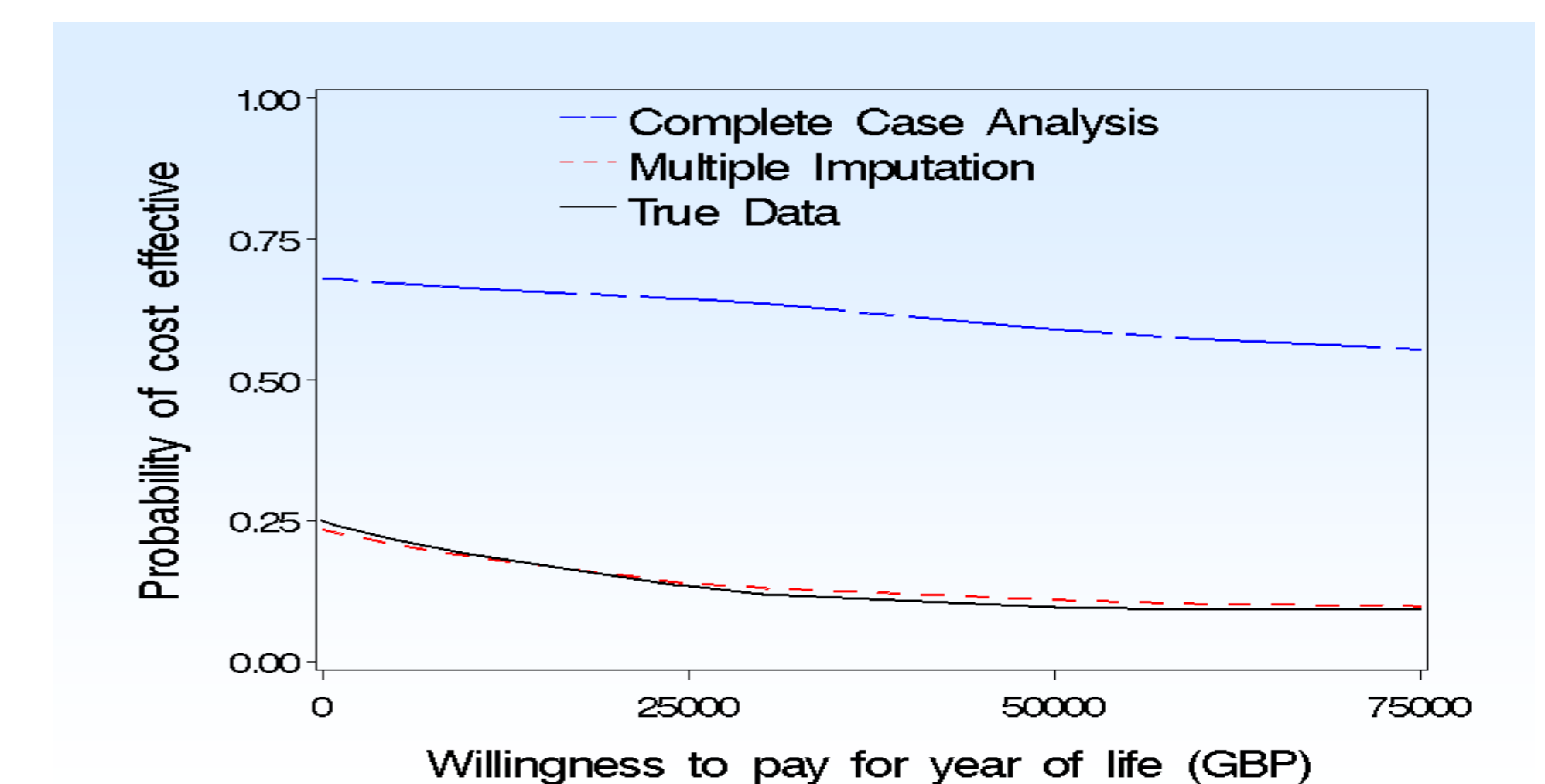
- Further cost-effectiveness analysis was performed by using 2000 treatment-stratified bootstrapping samples comparing treatment B to treatment A.
- True data: treatment A was both more effective (in terms of survival) and less expensive (considering the above 3 cost components) than treatment B.
- Scatter plot from MI data provided the same evidence as the true data that subjects allocated to the treatment arm B had a higher cost than those allocated to treatment A and the difference in mean survival time between treatment B and treatment A was negative while CCA analysis showed the opposite pattern.

Figure 3: Scatter plots of the difference in mean cost and effectiveness on a cost-effectiveness plane (treatment B vs. treatment A). Left - True data, Middle - Complete case analysis, Right - MI analysis.



- Taking a threshold of willingness to pay of £24,000 per life-year gained, net monetary benefit (NMB) confirmed different conclusion from CCA and MI analyses
- For CCA, the mean NMB for treatment B was £1664 (95% CI £1476 to £1853), suggesting that treatment B is cost-effective compared to treatment A
- After MI, the NMB for treatment B was -£3599 (95% CI -£3646 to -£3552), which implies that treatment B is not cost-effective compared to treatment A
- The cost-effectiveness acceptability curve (CEAC) for MI showed that treatment B has never been the more cost-effective treatment option for values of λ from £0 to £75,000 per life-year gained while for CCA it showed that probability of treatment B being cost-effective was always higher than 50%.

Figure 4: Cost-effectiveness acceptability curve



Discussion

Within health economic evaluations, complete case analysis remains the primary analysis, irrespective of the amount of missing data, although sensitivity analyses are sometimes performed to assess the effect of missing data.

By ignoring subjects with missing cost components CCA will lead to inefficient prediction. Multiple imputation technique is an easily implemented computer intensive method which, by including as many relevant covariates (to the missingness or to the cost components) in imputation model under the assumption of missing at random, largely enhances the efficiency of prediction. When missingness has a non-monotone pattern, MCMC should be applied first to impute enough data to make the missing pattern monotone. Since the parameter convergence in MCMC can be checked by time-series and autocorrelation plot and an informative prior can be also specified through the Bayes theorem, this method was recommended to impute the missing individual cost components for further cost-effectiveness analysis. MCMC technique has been adopted by many authors (e.g., Nicholas J. Horton, Stuart R. Lipsitz. [2001]; Jan B. Oostenbrink and Maiwenn J. [2005]) to fill in the missing cost data prior to performing a realistic cost-effectiveness analysis. Although the way to handle incomplete and censored cost information remains subject to debate, MI techniques such as regression and MCMC imputing missing costs at the patient level offer a valid alternative to classical analysis methods such as ACA and CCA and yield realistic results. Such models can include several contributing covariates including baseline variables, actual survival, as well as the costs from previous time points to improve the prediction.

CONCLUSION

- ACA tends to underestimate the total costs
- CCA leads to biased results as complete cases are not representative of the original patient population
- MI techniques are valuable to produce a complete cost datasets especially when the assumption of missing completely at random is questionable
- Regression and MCMC were preferred to propensity score, which is a non-parametric technique and overlooks the relationships between costs and covariates.